HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

ANCIENT DNA SCREENING FROM FINNISH STONE AGE SEDIMENTS

Sanni Peltola

Genetics and Molecular Biosciences

Faculty of Biological and Environmental Sciences

University of Helsinki

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildingsprogram – Degree Programme
Biological and environmental sciences	Genetics and molecular biosciences

Tekijä – Författare – Author

Sanni Peltola

Työn nimi – Arbetets titel – Title

Ancient DNA screening from Finnish Stone Age sediments

Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track

Genetics and genomics

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Master's thesis	September 2019	51

Tiivistelmä – Referat – Abstract

In recent decades, ancient DNA recovered from old and degraded samples, such as bones and fossils, has presented novel prospects in the fields of genetics, archaeology and anthropology. In Finland, ancient DNA research is constrained by the poor preservation of bones: they are quickly degraded by acidic soils, limiting the age of DNA that can be recovered from physical remains. However, some soil components can bind DNA and thus protect the molecules from degradation. Ancient DNA from soils and sediments has previously been used to reconstruct paleoenvironments, to study ancient parasites and diet and to demonstrate the presence of a species at a given site, even when there are no visible fossils present. In this pilot study, I explored the potential of archaeological sediments as an alternative source of ancient human DNA. I collected sediment samples from five Finnish Neolithic Stone Age (6,000–4,000 years ago) settlement sites, located in woodland. In addition, I analysed a lakebed sample from a submerged Mesolithic (10,000–7,000 years ago) settlement site, and a soil sample from an Iron Age burial with bones present to compare DNA yields between the two materials. Soil samples were converted into Illumina sequencing libraries and enriched for human mtDNA. I analysed the sequencing data with a customised metagenomics-based bioinformatic analysis workflow. I also tested program performance with simulated data. The results suggested that human DNA preservation in Finnish archaeological sediments may be poor or very localised. I detected small amounts of human mtDNA in three Stone Age woodland settlement sites and a submerged Mesolithic settlement site. One Stone Age sample exhibited terminal damage patterns suggestive of DNA decay, but the time of deposition is difficult to estimate. Interestingly, no human DNA was recovered from the Iron Age burial soil, suggesting that body decomposition may not serve as a significant source of sedimentary ancient DNA. Additional complications may arise from the high inhibitor content of the soil and the abundance of microbial and other non-human DNA present in environmental samples. In the future, a more refined sampling approach, such as targeting microscopic bone fragments, could be a strategy worth trialling.

Avainsanat – Nyckelord – Keywords

ancient DNA, metagenomics, archaeogenetics, sediments, bioinformatics, Stone Age, environmental DNA

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors

Päivi Onkamo

Säilytyspaikka – Förvaringställe – Where deposited

Ethesis

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildingsprogram – Degree Programme
Bio- ja ympäristotieteellinen	Genetiikka ja molekulaariset biotieteet

Tekijä – Författare – Author

Sanni Peltola

Työn nimi – Arbetets titel – Title

Muinais-DNA:n säilyvyys Suomen kivikautisissa sedimenteissä

Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track

Genetiikka ja genomiikka

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Pro gradu -tutkielma	Syyskuu 2019	51

Tiivistelmä – Referat – Abstract

Viime vuosikymmeninä vanhoista ja hajonneista näytteistä eristetty muinais-DNA on avannut uusia mahdollisuuksia niin genetiikan, arkeologian kuin antropologiankin tutkimuksessa. Suomessa muinais-DNA-tutkimusta rajoittaa luiden nopea hajoaminen happamassa maaperässä, jolloin luumateriaalia käyttäen on mahdollista tutkia vain melko viimeaikaisia ajanjaksoja. DNA sitoutuu kuitenkin myös joihinkin maaperän rakenneosiin, jolloin sen hajoaminen hidastuu ja se voi säilyä maaperässä, vaikka näkyviä jäänteitä ei enää ole. Aiemmin maaperästä tai sedimenteistä eristettyä muinaista DNA:ta on käytetty muinaisten ekosysteemien ja muinaisten ihmisyhteisöjen loisten ja ruokavalion tutkimiseen sekä osoittamaan esihistoriallisten lajien läsnäolo tutkitulla paikalla. Tässä pilottitutkimuksessa kartoitin Suomen arkeologisten sedimenttien käyttöä muinaisen ihmis-DNA:n vaihtoehtoisena lähteenä. Tutkimusta varten keräsin sedimenttinäytteitä viideltä kivikautiselta (6000–4000 vuotta sitten) asuinpaikalta. Lisäksi analysoin sedimenttinäytteitä mesoliittiselta (10000-7000 vuotta sitten), veden alle jääneeltä asuinpaikalta, sekä maanäytteen luuainesta sisältäneestä rautakautisesta ruumishautauksesta verratakseni DNA:n määrää luussa ja maaperässä. Maanäytteistä valmistettiin Illumina-sekvensointikirjastot ja niistä rikastettiin ihmisen mitokondrio-DNA:ta. Analysoin sekvenssit metagenomiselle aineistolle suunnitellulla menetelmällä, jonka suorituskykyä testasin myös simuloidulla aineistolla. Tulosten perusteella ihmis-DNA:n säilyy Suomen maaperässä huonosti tai hyvin paikallisesti. Havaitsin pieniä määriä ihmisen mtDNA:ta kolmessa kivikautisesta ja yhdessä mesoliittisessa kohteessa. Yhdessä kivikautisessa näytteessä oli merkkejä DNA:n hajoamisesta, mutta eristettyjen molekyylien ikää on vaikea arvioida. Rautakautisen ruumishaudan maaperästä ei löytynyt lainkaan ihmis-DNA:ta, mikä viittaa siihen, ettei ruumiin hajoamisesta jää merkittäviä määriä DNA:ta maaperään. Arkeologisten sedimenttien käyttöä muinais-DNA:n lähteinä saattavat lisäksi rajoittaa maaperän inhibiittorit sekä ympäristö-DNA:n valtava määrä suhteessa kohde-DNA:han. Jatkossa näytteentoton huolellisempi kohdentaminen saattaisi osittain vastata näihin haasteisiin.

Avainsanat – Nyckelord – Keywords

Muinais-DNA, metagenomiikka, arkeogenetiikka, sedimentit, bioinformatiikka, kivikausi, ympäristö-DNA

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors

Päivi Onkamo

Säilytyspaikka – Förvaringställe – Where deposited

Ethesis

Muita tietoja – Övriga uppgifter – Additional information

Table of contents

1 INTRODUCTION 2

2 AIMS 7

3 MATERIALS AND METHODS 7

3.1 Sample collection 7

3.2 Laboratory settings 10

3.3 DNA extraction 11

3.4 DNA library preparation 13

3.5 Human mitochondrial enrichment 19

3.6 Sequencing 22

3.7 Validation of the bioinformatic analysis workflow 22

3.8 Analysis of the sequencing data 25

4 RESULTS 29

4.1 Human mtDNA recovery and ancient DNA authenticity 29

4.2 Kraken: results from simulated data 31

4.3 Inhibitor carry-over in DNA extraction 35

4.4 Library quantification and amplification efficiency 36

4.5 Data quality and contamination 39

5 DISCUSSION 41

6 CONCLUSIONS 46

Acknowledgements 47

References 48

Appendices i–viii

Abbreviations

aDNA	Ancient DNA
mtDNA	Mitochondrial DNA
WGS84	The World Geodetic System standard, 1984 revision
CE	Common era
NaClO	Sodium hypochlorite; bleach
IEM	Institute of Evolutionary Medicine, University of Zürich
UV	Ultraviolet
PCR	Polymerase chain reaction
EDTA	Ethylenediaminetetraacetic acid
rpm	Rounds per minute
g	gravitational force equivalent
PE buffer	Qiagen's commercial wash buffer for DNA cleanup procedures
TE buffer	DNA solution buffer
qPCR	Quantitative PCR
T4 PNK	T4 polynucleotide kinase
dNTP	deoxyribonucleotide triphosphate
ATP	Adenosine triphosphate
PB buffer	Qiagen's commercial binding buffer for DNA cleanup procedures
TET buffer	TE buffer with Tween-20
Bst polymerase	A heat-resistant polymerase with strand displacement activity; originally from *Bacillus stearothermophilus*
Pfu polymerase	DNA polymerase, originally from *Pyrococcus furiosus*
Cq	The qPCR cycle number at which the fluorescent signal intersects the detection threshold.
Hi-RPM buffer	Agilent's commercial hybridization buffer
BWT	Bind&Wash&Tween buffer
HWT	Hybridization wash buffer
NCBI	National center for biotechnology information
k-mer	A *k* nucleotides long subsequence of a longer sequence
LCA	Lowest common ancestor of a group of sequences
RTL-path	Root-to-leaf path; a path leading from a higher to lower taxonomic level in a phylogeny.
EAGER	Efficient Ancient Genome Reconstruction; an analysis pipeline for aDNA
BWA	Burrows-Wheeler Aligner

1 INTRODUCTION

The past two decades have witnessed immense advancements in ancient DNA (aDNA) research, facilitated by developments in next generation sequencing technology. These innovations, together with refinements in laboratory protocols and computational methods, now allow the retrieval and sequencing of DNA of deceased organisms that can be up to hundreds of thousands of years old (Meyer et al., 2014). Ancient DNA can be utilised in a variety of research topics: for example, it can help to resolve the evolutionary history of extinct species, shed light on past population events and unravel domestication processes and historical plagues. Furthermore, sequencing methods developed for ancient DNA have applications in other fields where retrieving DNA from degraded samples may be of interest, such as forensics and medicine (Overballe-Petersen et al., 2012). In addition to fossils and archaeological finds, aDNA can been retrieved from various organic materials, such as museum specimens, formalin-fixed medical samples, coprolites and sediments (Gansauge and Meyer, 2013; Hofreiter et al., 2001; Pääbo et al., 2004; Willerslev et al., 2003).

A window to our genetic past

To date, human remains have been the most popular source of ancient DNA, and the results from these studies have inevitably changed our understanding of the distant and recent history of our species. Ancient DNA from Pleistocene hominin fossils has convincingly shown that modern humans and Neanderthals interbred, and suggested that interbreeding was common whenever two hominin species met (Green et al., 2010; Slon et al., 2018; Reich et al., 2011; Fu et al., 2015). The discovery of the Denisovans, a sister group of the Neanderthals, was almost solely based on ancient DNA, as DNA from a tiny finger bone from a Siberian cave turned out to belong to this previously unknown hominin species (Krause et al., 2010). Through ancient DNA, it is possible not only to understand the biology and demography of these long-lost species, but also to gain insights into our own evolutionary history. For example, by comparing the genomes of extinct hominins and modern humans it is possible to track down putatively adaptive traits that are unique for our species (Sanchez-Quinto and Lalueza-Fox, 2015).

The first attempts to retrieve and analyse ancient DNA from fossils and archaeological samples were tedious tasks. Luckily, these early studies have facilitated standardised procedures and analysis software that now make ancient DNA studies much faster and easier. With denser sampling and refined analysis, it has become possible to study prehistoric and historic population events in

most parts of the world. For instance, large-scale studies on human remains from Europe have given us a detailed picture of the population structure and migrations within the past 10,000 years. We have learned that European Mesolithic hunter-gatherers were genetically distinct from succeeding early Neolithic farmers, and that this turnover in genetic makeup was associated with the spread of agriculture from the Near East and Anatolia approximately 8,000–5,000 years ago (Haak et al., 2010; Skoglund et al., 2012; Gamba et al., 2012). Later, at the advent of the Bronze Age, another massive gene flow from the East Eurasian steppe left its signature on the genetic composition of modern Europeans (Lazaridis et al., 2014; Allentoft et al., 2015; Haak et al., 2015).

Whereas most modern European populations can be modelled as mixtures of these three genetic components – Mesolithic hunter-gatherer, Neolithic early farmer and Bronze Age steppe – some populations require a fourth component to explain their genetic legacy (Lazaridis et al., 2014). Finns and Saami are among these outlier populations, and carry additional, Siberian-related ancestry (Lamnidis et al., 2018). Despite the peculiar population history of the region, northern and northeastern parts of Europe remain understudied.

DNA degradation and ancient DNA damage

Major challenges in ancient DNA studies come from the natural degradation of biomolecules. After organisms die, their DNA starts to decay as the cellular machinery that normally maintains DNA integrity ceases to operate (Lindhal, 1993). Various factors, most notably endogenous and environmental microbes and intracellular nucleases released from their cellular compartments contribute to the immediate post-mortem degradation of DNA (Pääbo et al., 2004). Favourable conditions, such as cold temperature or desiccation, can significantly slow down DNA decay. However, even under ideal environmental conditions, DNA does not preserve forever. Oxidative and hydrolytic damage will continue to accumulate until no retrievable sequence information remains (Dabney et al., 2013a). The limit of DNA preservation is not completely clear, but it is thought to be around one million years. Currently, the oldest DNA has been recovered from Greenland ice cores and is approximately 800,000 years old (Willerslev et al., 2007). Outside permafrost DNA preservation is more limited and highly dependent on environmental factors, such as humidity, temperature, salinity and pH, which can vary significantly even within a single site (Dabney et al., 2013a).

Even in the most well-preserved samples, endogenous DNA – the DNA of the organism itself – is present only in trace amounts and in a degraded state. Most DNA molecules in ancient samples originate from microbes that took over and degraded the remains after the death of the organism, while endogenous DNA makes up only a minor portion (Green et al., 2010). Moreover, due to DNA decay, most endogenous DNA molecules are very short and extensively damaged. Hydrolytic and oxidative processes are the major forces that break down DNA over time (Lindhal, 1993). Hydrolytic cleavage of purine residues leads to abasic sites, which are susceptible to single-strand breakage and subsequent DNA fragmentation (Overballe-Petersen et al., 2012; Briggs et al., 2007). Hydrolysis is also responsible for cytosine deamination, a major form of miscoding lesions typical in ancient DNA (Hofreiter et al., 2001). The loss of an amino group from cytosine produces uracil, which is read as thymine by DNA polymerase, and deaminated cytosines appear as C-to-T substitution in subsequent sequence analyses (Figure 1). Cytosine deamination mostly affects the ends of a DNA fragment, suggesting that it mainly occurs on the single-stranded overhangs of a fragmented double-stranded molecule (Briggs et al., 2007).

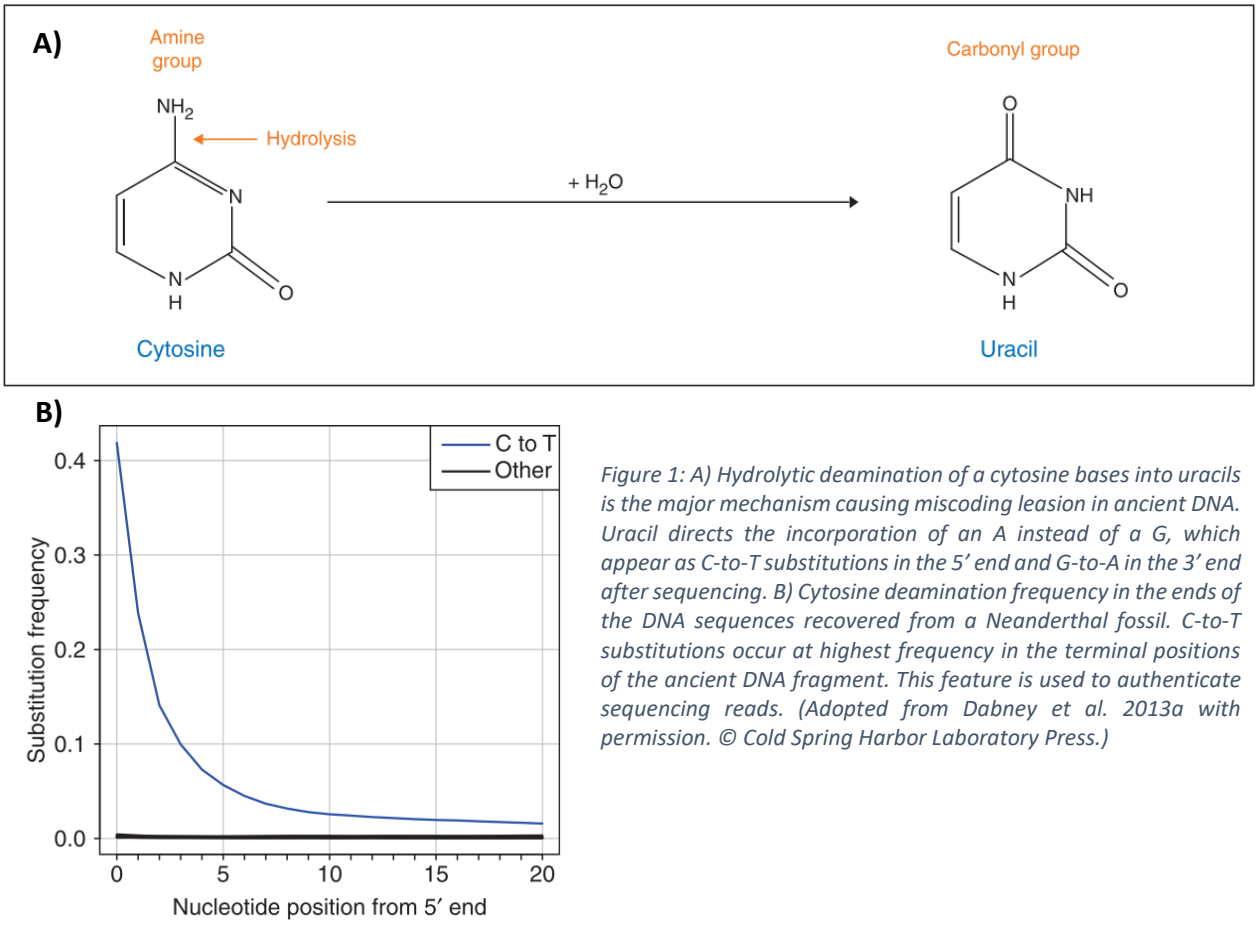

Figure 1: A) Hydrolytic deamination of a cytosine bases into uracils is the major mechanism causing miscoding leasion in ancient DNA. Uracil directs the incorporation of an A instead of a G, which appear as C-to-T substitutions in the 5' end and G-to-A in the 3' end after sequencing. B) Cytosine deamination frequency in the ends of the DNA sequences recovered from a Neanderthal fossil. C-to-T substitutions occur at highest frequency in the terminal positions of the ancient DNA fragment. This feature is used to authenticate sequencing reads. (Adopted from Dabney et al. 2013a with permission. © Cold Spring Harbor Laboratory Press.)

Oxidative processes cause other forms of damage such as bulky lesions that impede DNA polymerases (Dabney et al., 2013a). Polymerase blocking lesions are less well characterised than strand break causing damage or nucleotide damage, but according to some studies, they can be present at a noticeably high frequency in ancient DNA and thus severely hinder DNA recovery.

The small amount and the damaged nature of the molecules makes ancient DNA sequencing prone to contamination. Therefore, ancient DNA research is carried out with several precautions to protect the samples from modern DNA (Fulton and Shapiro, 2019). Standard procedures include dedicated laboratories and special computational methods to distinguish modern contamination from authentically ancient DNA. For instance, benefit is derived from the ancient DNA damage: cytosine deaminations in the terminal positions of the DNA fragments are used as an indicator of DNA authenticity (Dabney et al., 2013a; Skoglund et al., 2014).

Ancient DNA from soils and sediments

Recently, the first ancient DNA results from mainland Finland have been published (Lamnidis et al., 2018). The oldest samples in the study are from the water burial of Levänluhta and date back to the Finnish Iron Age (300–800 CE). Bone preservation in Finland is generally poor, and physical remains from earlier human occupation are rare or absent (Ahola et al., 2016). Due to acidity, unburned bones tend to degrade in less than 2,000 years, restricting the achievable time depth for aDNA studies. Consequently, it seems difficult to trace back putative changes in the genetic makeup of Finland before the Iron Age. To overcome these limitations, alternative sources of ancient human DNA have been considered: for example, ancient DNA was recently recovered from Swedish Stone Age birch bark pitch mastics (Kashuba et al., 2019), which are also abundant in Finnish archaeological record. In addition, archaeological sediments hold potential as a source of ancient DNA. Compared to bones or other physical human remains, sediments are abundant and often easily accessible. Sediment samples are easier and faster to prepare for DNA extraction, since there is no need to drill or homogenise the sample material. There are also fewer ethical considerations, since soil sampling can be done without damaging valuable and rare archaeological items, and there is no handling of human remains.

DNA survival in soil depends on various physical, chemical and biological factors. Extracellular DNA released into soil is vulnerable to degradation by microbial DNases but binding on charged organic and mineral soil components can significantly prolong its survival (Pedersen et al., 2014).

Clay minerals in particular can bind extensive amounts of DNA. Sand minerals, such as quartz, which are more common in Finnish soils, can also bind DNA, but to a lesser extent (Ogram et al., 1994; Slon et al., 2017). Humic acids, which make up the major organic soil component, are also known to bind DNA, although their presence in extracts may inhibit DNA polymerase activity and hamper downstream analyses (Crecchio and Stotzky, 1998; Rohland et al., 2018).

Unlike with physical remains, the source of ancient DNA in sediments is often unclear. Human or animal DNA bound to soil components may originate from a number of sources, such as excrement, blood, or body decomposition (Pedersen et al., 2014; Slon et al., 2017). Ancient DNA from various organisms can also be present in microscopic bone fragments, coprolite pieces, parasite eggs, or plant seeds incorporated into sediment. The authentication and dating of sequences retrieved from ancient sediments can be more challenging compared to sequences retrieved from physical remains, such as bones: DNA from the soil surface can sometimes leach through strata and contaminate the original layer of interest (Haile et al., 2007). Additionally, redeposition, i.e. the mixing of strata, can also disturb the original context of the sample and make it difficult to reliably date. If sediment contains microfossils, it is possible to target them e.g. by fixing the undisturbed sediment blocks with chemical resins (Massilani et al., 2018). This approach can both simultaneously increase DNA recovery and help authenticate the results.

Ancient DNA has been successfully retrieved from basal ice cores, lake cores, and surface soils and sediments. These materials have provided a rich source of ancient environmental DNA for paleoecological studies. For example, plant and insect DNA recovered from Greenland basal ice cores provided evidence that the area, which is now two kilometres below ice, was once forested (Willerslev et al., 2007). In Alaska, ancient DNA from perennially frozen sediments revealed that mammoths and horses had survived in north-western North America at least until 10,500 years ago, several thousand years longer than previously thought (Haile et al., 2009). For paleoecological studies, DNA from sediments is most useful when combined with complementary information from other sources, such as fossil records or pollen analyses (Pedersen et al., 2014).

Although most studies on sedimentary ancient DNA have so far focused on non-human organisms, archaeological sediments can be applied to gain insights into human history. For instance, ancient sediments from latrines, waste pits, and submerged settlement sites have been studied to identify parasite species and dietary components present in prehistorical communities (Søe et al., 2018; Tams et al., 2018; Smith et al., 2015). Results can be used to infer changes in

subsistence, as well as ancient health and diet. However, human DNA can be directly retrieved from sediments as well, at least under some conditions. Slon et al. (2017) have demonstrated that tens of thousands of years old Neanderthal and Denisovan DNA persists in Pleistocene cave sediments. They retrieved ancient mitochondrial DNA of several mammalian taxa that had been present in the cave during the Pleistocene. These species also included extinct hominins, and one of the sediment samples contained enough DNA to reconstruct a full mitochondrial genome. Although their results only included mitochondrial DNA, which is usually preserved in higher numbers than nuclear DNA due to its high copy number in cells, the level of DNA preservation implies that retrieving nuclear DNA from cave sediments is not out of reach. Therefore, sediments could have potential as an additional source of ancient human DNA, especially in regions where human remains are sparse.

2 AIMS

The aim of this study was to explore prospects of ancient human DNA preservation in Finnish archaeological sediments and to assess their potential applications for archaeogenetic research in Finland. The main target was human mitochondrial DNA, but the sequencing data was also screened for an array of other mammalian taxa, whose presence in archaeological settlement sites could provide information on prehistoric animal consumption. Additionally, my goal was to customise an existing bioinformatic analysis workflow for ancient environmental sequencing data.

3 MATERIALS AND METHODS

3.1 Sample collection

Neolithic Stone Age settlement sites

Five Stone Age settlement sites were sampled specifically for this project. The site selection and sampling excursion was done together with professional archaeologists. We collected sediment samples from three sites located near the coast of the Gulf of Finland (Loviisa Spångkärret, Kotka Niskasuo and Virolahti Karpankangas), and two from inland South Karelia (Taipalsaari Konstunkangas and Taipaleenranta 2) (Pesonen, 2018; Table 1; Figure 2). All settlement sites are currently woodland, and the soil of the sites varies from rough sandy gravel to fine sand. The chosen sites have not been fully excavated, and thus remain mostly undisturbed. DNA analysis on the sediments was considered as an easy, relatively non-invasive, and affordable way to further study these sites. Since archaeological excavations are always a possible source of contamination, the lack

of thorough excavations can be seen as a benefit. Finds from the sites include mainly comb ceramics, quartz, and pieces of burnt bone, which were also encountered during the sampling. Ceramic finds from Spångkärret, Niskasuo, Karpankangas and Konstunkangas represent middle Neolithic Stone Age cultures, while ceramics from Taipalsaari 2 are from the later Neolithic Stone Age. In Finland, middle Neolithic dates to 5,900–5,200 years ago and later Neolithic to 5,250–4,500 years ago (Haggren et al., 2015), making Taipalsaari 2 slightly younger than the rest of the sites.

The chosen sites are characterised by "house pits": shallow, 5–10 meters wide depressions in the ground left by prehistoric dwellings. Most sites have several house pits located close to each other. Sampling was partly concentrated onto the inner bank of the pits, because we reasoned that possible DNA sources, such as human excrements and other waste materials, might have accumulated in those areas when people lived in the settlements. We chose two house pits from each site for sampling and collected 2–4 samples from each pit. In Taipaleenranta 2, we could only locate one house pit, where we collected five samples.

Sampling was done with a small hand-held earth drill. Prior to the sampling of each house pit, we rinsed the drill with 10% Bleach (NaClO) to destroy modern DNA from the tool's surface, let it incubate for 10–15 minutes, and rinsed it with sterile water to remove excess bleach. A sediment block was brought to the surface with the drill, and a sample was collected from the dark "culture horizon" of the sediment block whenever visible, approximately from 20–30 cm depth. The surface of the block was first carefully removed, and a small amount of soil was transferred into a clean, UV-treated 15-mL Falcon tube with a sterile scalpel. All participants wore respiration masks and latex gloves during bleaching, drilling and sampling to avoid modern contamination. Due to the high frequency of unsuccessful drilling attempts and limited time, we decided not to repeat bleaching between individual drillings, although we recognised that this could predispose our samples to cross contamination.

Table 1: Names, coordinates and IDs of Stone Age settlement sites that were sampled for the purpose of this project.

MUNICIPALITY	SITE	LATITUDE (WGS84)	LONGITUDE (WGS84)	FINNISH HERITAGE AGENCY SITE ID
LOVIISA	Spångkärret	60.52253703°	26.24989437°	1000027878
KOTKA	Niskasuo	60.58753772°	26.79757473°	285010017
VIROLAHTI	Karpankangas	60.60998560°	27.78705967°	1000014201
TAIPALSAARI	Konstunkangas	61.13435938°	28.07296267°	831010005
TAIPALSAARI	Taipaleenranta 2	61.15998857°	28.09594838°	831010029

In total, 32 samples were collected from the five sites during a two-day trip in July 2018, after which the samples were stored in a freezer. We chose 12 samples for subsequent aDNA screening: two from Loviisa Spågkärret, Kotka Niskasuo, and Virolahti Karpankangas, and three from Taipalsaari Konstunkangas and Taipaleenranta 2.

Submerged Mesolithic settlement site

One additional Stone Age sample was acquired from the Lost Inland Landscapes project, led by Satu Koivisto. A submerged Mesolithic settlement site was discovered from the bottom of Kammarlahti, Lake Kuolimo, during underwater excavations in the summer 2018 (Figure 2). A sediment sample was collected from a fireplace that was discovered during the excavations. Because the sample was originally not collected for the purpose of ancient DNA analysis, contamination was not considered at the time, but the sample was kept cold and undisturbed after collection. I took subsamples for genetic analyses from the sediment sample in August 2018 following precautions to prevent modern contamination. The subsampling was done in the Finnish Heritage Agency's Collections and Conservation Centre, where no PCR has ever been done, wearing latex gloves and respiration filters. The workspace was cleaned with 10% NaClO before handling the sample. The surface of the soil block was removed, and each subsample was transferred into a sterile 15-mL Falcon tube with a disposable sterile scalpel. The subsamples were then stored in a freezer. Five subsamples were collected and three of these were selected to be part of the ancient DNA analysis.

Iron Age burial

To compare the extent of DNA preservation between soil and bone, I included a soil sample from a Late Iron Age (800–1200 CE) burial from Valkeakoski Toppolanmäki (Figure 2). The soil type of the burial site is similar to Neolithic settlement sites and consists of fine-grained sand. Ancient DNA has been successfully extracted and sequenced from the bone material of the grave, confirming that DNA preservation in the microenvironment of that burial is possible (unpublished data).

The sediment samples were collected in June 2017 from grave 3 from Valkeakoski Toppolanmäki cemetery (Moilanen, 2017). The grave had been previously opened and refilled during excavations in 1937. The condition of the bones in 2017 suggested that the exhumation 70 years earlier had had very little impact on the bone preservation. The grave was approximately 65 cm deep, but the pelvis, sacrum and lower vertebrae lay deeper than the skull and remaining long bones and had not been

revealed in the 1937 excavation. Since ancient DNA analyses were planned, precautions were taken to avoid contamination were taken when opening the grave and collecting the samples. In addition to bones, four soil samples were collected from the pelvic area of the individual and stored in a freezer.

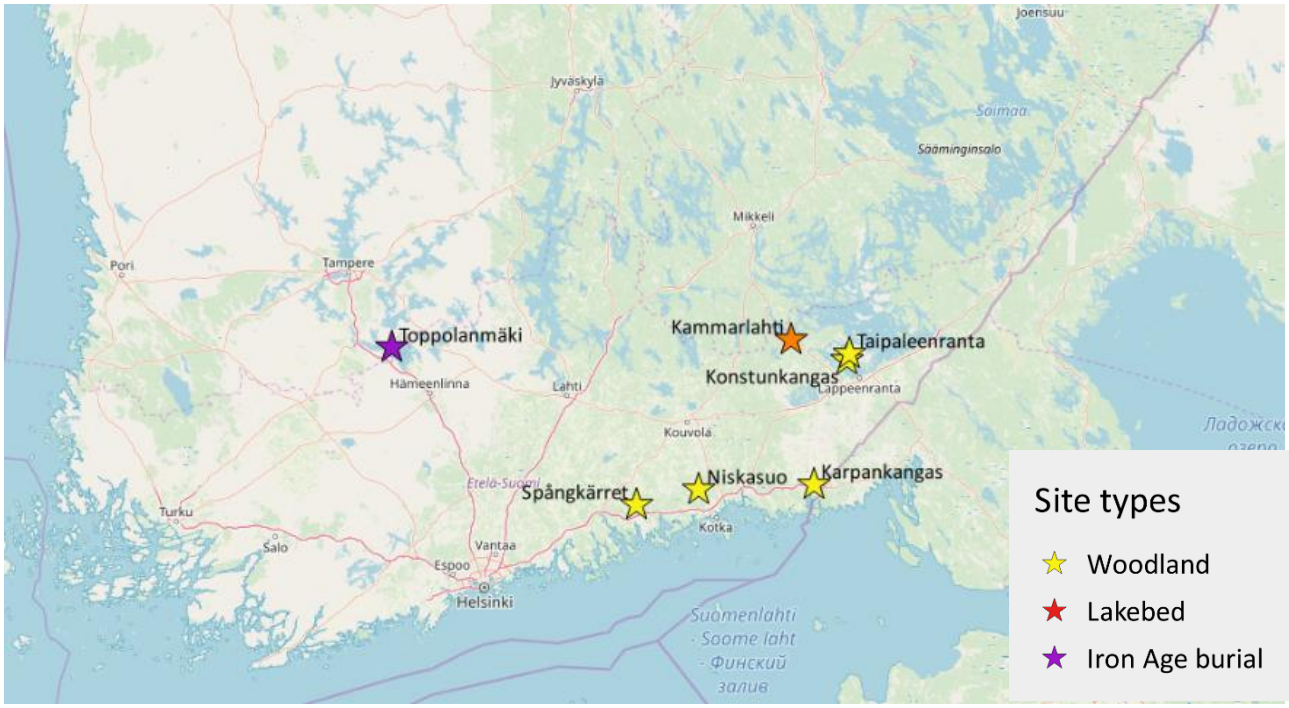

Figure 2: A map with archaeological sites included in the project.

3.2 Laboratory settings

The laboratory work was carried out in the ancient DNA laboratory of the Institute of Evolutionary Medicine, University of Zürich, Switzerland in collaboration with the local Paleogenetics group led by Verena Schünemann. The facilities have been designed for ancient DNA studies, and strict precautions to prevent contamination from modern DNA were followed. All contamination-prone pre-amplification steps were carried out in a "clean room", a special laboratory designed for ancient DNA research. The clean room has positive air pressure to prevent airflow from the outside of the room, and surfaces are lit with UV lamps to destroy lingering DNA when the laboratory is not in use. The clean room is divided into four separate spaces for different work steps, and there are separate laminar flow hoods for samples and reagents to prevent cross-contamination between samples. All surfaces are cleaned with ultrapure water and DNA Away (Thermo Scientific) after use or after any spilling of liquids. Labware is cleaned with bleach. Since the most obvious source of contamination are the people who work in the laboratory, the clean room can only be entered wearing shoe covers, a full-body Tyvek suit, respiration filter, hairnet, faceguard, and three layers

of gloves. The outermost gloves are changed regularly while working. Only ultrapure water is used for both reactions and cleaning.

PCR amplifications and post-PCR steps are carried out in a separate "modern lab", following normal laboratory precautions. As per normal precaution operations, reagents or samples must not be transferred from post-PCR laboratory to the clean room, and people are not allowed into the clean room after working in the post-PCR laboratory during the same day.

3.3 DNA extraction

To extract DNA from the archaeological sediment samples, we followed the protocol presented by Dabney et al., 2013b, which is optimised for ancient DNA and should efficiently recover short DNA fragments. It also removes inhibitors more efficiently than other ancient DNA extraction methods, and is therefore the recommended protocol for ancient sediments (Rohland et al., 2018). The desired fragment length distribution is achieved with a special binding buffer composition and using an increased binding volume (Dabney et al., 2013b). Extracted DNA is purified by binding it to silica membrane.

From each sediment sample, 45–100 μg of sediment was measured into a 2-mL Eppendorf tube (Table 2). Two empty 2-mL tubes for negative extraction controls were labelled, and they were carried alongside the samples through all steps to monitor laboratory contamination. For each sample, one plastic extension reservoir was prepared. We used reservoirs removed from commercial Zymo-Spin V columns (Zymo Research). Reservoirs were removed from the columns, the columns were discarded, and the reservoirs were cleaned by soaking them in 7% bleach, rinsing twice in water and UV irradiating in a cross-linker for 30 minutes. In addition, one 50-mL and one 15-mL Falcon tube was UV-irradiated in the cross-linker for 30 minutes for each sample. Extraction buffer with 0.45 M EDTA (Fischer scientific) and 0.25 mg/mL proteinase K (BioConcept AG) was prepared by mixing 17.1 mL of 0.5 M EDTA, 1425 μL of water and 475 μL of 10 mg/mL proteinase K; proteinase K was added right before adding the extraction buffer to the samples. One mL of extraction buffer was added to each sediment sample, and samples were sealed with parafilm and incubated overnight in 37 °C on rotation.

Table 2: List of sediment samples that were subsampled for DNA extraction, subsample laboratory ID and weight.

SITE	SAMPLE NAME	SAMPLE LABID	SAMPLE WEIGHT (G)
LOVIISA SPÅNGKÄRRET	SPÅ P1 1/3	ZH0724	0.55
	SPÅ P2 3/3	ZH0725	0.57
KOTKA NISKASUO	NISKA P1 1/4	ZH0726	0.53
	NISKA P2 2/2	ZH0727	0.49
VIROLAHTI KARPANKANGAS	KAR P1 1/3	ZH0728	0.47
	KAR P2 2/4	ZH0729	0.50
TAIPALSAARI KONSTUNKANGAS	KONSTU P1 2/4	ZH0730	0.62
	KONSTU P2 2/4	ZH0731	0.58
	KONSTU P2 4/4	ZH0732	0.52
TAIPALSAARI TAIPALEENRANTA	TAI 2/5	ZH0733	0.58
	TAI 3/5	ZH0734	0.51
	TAI 4/5	ZH0735	0.68
SAVITAIPALE KAMMARLAHTI	1/5	ZH0736	0.62
	3/5	ZH0737	0.50
	4/5	ZH0738	0.56
VALKEAKOSKI TOPPOLANMÄKI	-	ZH0739	0.53
NEGATIVE CONTROLS	1. Extraction control	ZH0739EB1	
	2. Extraction control	ZH0739EB2	
	DNA Library preparation control	ZH0739LB	

For DNA extract purification, MinElute silica spin-columns (Qiagen) were used. The columns were removed from their collection tubes and attached to the cleaned and UV-treated extension reservoirs. Each reservoir–column complex was then placed into a 50-mL UV-irradiated Falcon tube to make a custom binding apparatus that allows a larger binding buffer volume to be used. The binding buffer containing 5 M guanidine hydrochloride (Sigma-Aldrich), 40% isopropanol (Sigma-Aldrich) and 90 mM sodium acetate (Sigma-Aldrich) was prepared by dissolving 119.4 g of guanidine hydrochloride in 150 mL of water and adding 100 mL of isopropanol. The mixture was aliquoted into UV-irradiated 15-mL Falcon tubes, where 10 mL of binding buffer mixture and 400 µL of 3 M sodium acetate was combined.

Incubated samples were centrifuged at 14,000 rpm (18,400 x g) in a tabletop centrifuge for 5 minutes to fix the solid particles to the bottom of the tube. Supernatant was carefully transferred into the binding buffer solution, either by pouring or pipetting, depending on the stability of the pellet. The mixtures were homogenised by mixing gently, and then poured into the extension

reservoir-column complexes placed inside the 50-mL Falcon tubes. the caps were closed, and the binding apparatuses were centrifuged at 1,500-2,800 rpm in 6-minute intervals until the whole extraction volume had passed through the silica membrane. The reservoir-column complexes were then removed from the Falcon tubes and placed back into 2-mL MinElute spin-column collection tubes. The extension reservoirs were detached from the columns and discarded. The columns were centrifuged at 6,000 rpm (3,380 x g) for 30 seconds to remove all the remaining liquid from them. Some samples were spun 2–3 times, because their flow-through was slower. After that, 700 µL of PE buffer (Qiagen) was added to each column and the columns were centrifuged at 6000 rpm for 30 seconds. Flow-through was discarded and the PE buffer wash was repeated. Centrifugations were repeated until all the liquid had passed through the silica membrane. The empty columns were then centrifuged twice more at 14,000 rpm for one minute to remove excess liquid; columns were rotated 180 degrees between centrifugations to ensure that no drops lingered in the corners. After that, the columns were transferred to clean 1.5 mL Eppendorf tubes. To elute DNA from the silica membrane, 50 µL of TE buffer (Sigma-Aldrich) was added to each column by carefully pipetting it onto the centre of the silica membrane. Columns were incubated for 5 minutes at room temperature and centrifuged at 13,000 rpm for one minute. Another 50 µL of TE buffer was added, and incubation and centrifugation were repeated, giving 100 µL of eluate in total. The columns were discarded, and the purified extracts were stored in a freezer.

3.4 DNA library preparation

High-throughput sequencing usually requires that extracted DNA molecules are converted into immortalised DNA libraries. In library preparation, adapters – artificial short sequences – are added adjacent to the ends of the extracted DNA molecules. Adapters contain priming sites that are used in amplification and sequencing (Meyer and Kircher, 2010). In addition, indices – short sample-specific oligonucleotides – can be incorporated into adapters: when several samples are sequenced in parallel, indices help to distinguish the origin of the sequencing read. However, there are several sources of errors that can cause a read to have an incorrect index, most notably "index bleeding" into neighbouring clusters during sequencing (Kircher et al., 2012). This phenomenon can be particularly pronounced with ancient DNA, where the DNA quantity and quality may differ between samples. Samples with the lowest DNA quantity are the most susceptible to misassignments (van der Valk et al., 2019). To avoid assigning reads to the wrong samples, two individually unique indices are used (Kircher et al., 2012). Indices are incorporated into adapters in

both ends of the library molecule and sequenced separately after the insert using a second primer set. Reads that carry unexpected index combinations can then be computationally removed after sequencing.

In ancient DNA research, both double-stranded and single-stranded libraries are used. Single-stranded library preparation has been found to improve sequence recovery when preparing damaged DNA fragments for sequencing, and it is often the recommended approach for the most ancient and degraded samples (Gansauge and Meyer, 2013). It can partially recover molecules that would be completely lost in double-stranded library preparation, such as molecules where both strands carry a single-strand breakage or where one of the stands has a polymerase-blocking lesion. However, it is more laborious and costly than double-stranded library preparation, and for the simple purpose of mtDNA screening, the double-stranded method was deemed sufficient.

Here, double-stranded and double-indexed DNA libraries were generated from the sediment sample DNA extracts and negative extraction controls (Meyer and Kircher, 2010; Kircher et al., 2012; Figure 3). One negative library control was added to monitor contamination during the library preparation. First, single-stranded overhangs in the extracted DNA molecules were repaired to form blunt ends, where adapters were subsequently ligated. Gaps in the 3'-ends were repaired in the fill-in reaction. Two sample-specific indices were introduced into adapters, and the full-length adapters were synthesised by amplification. The libraries were quantified with real-time quantitative PCR (qPCR) before and after indexing to monitor library preparation efficiency. Finally, the libraries were reamplified with Herculase II polymerase (Agilent) to reach an adequate library molecule copy number for shotgun sequencing and mitochondrial enrichment.

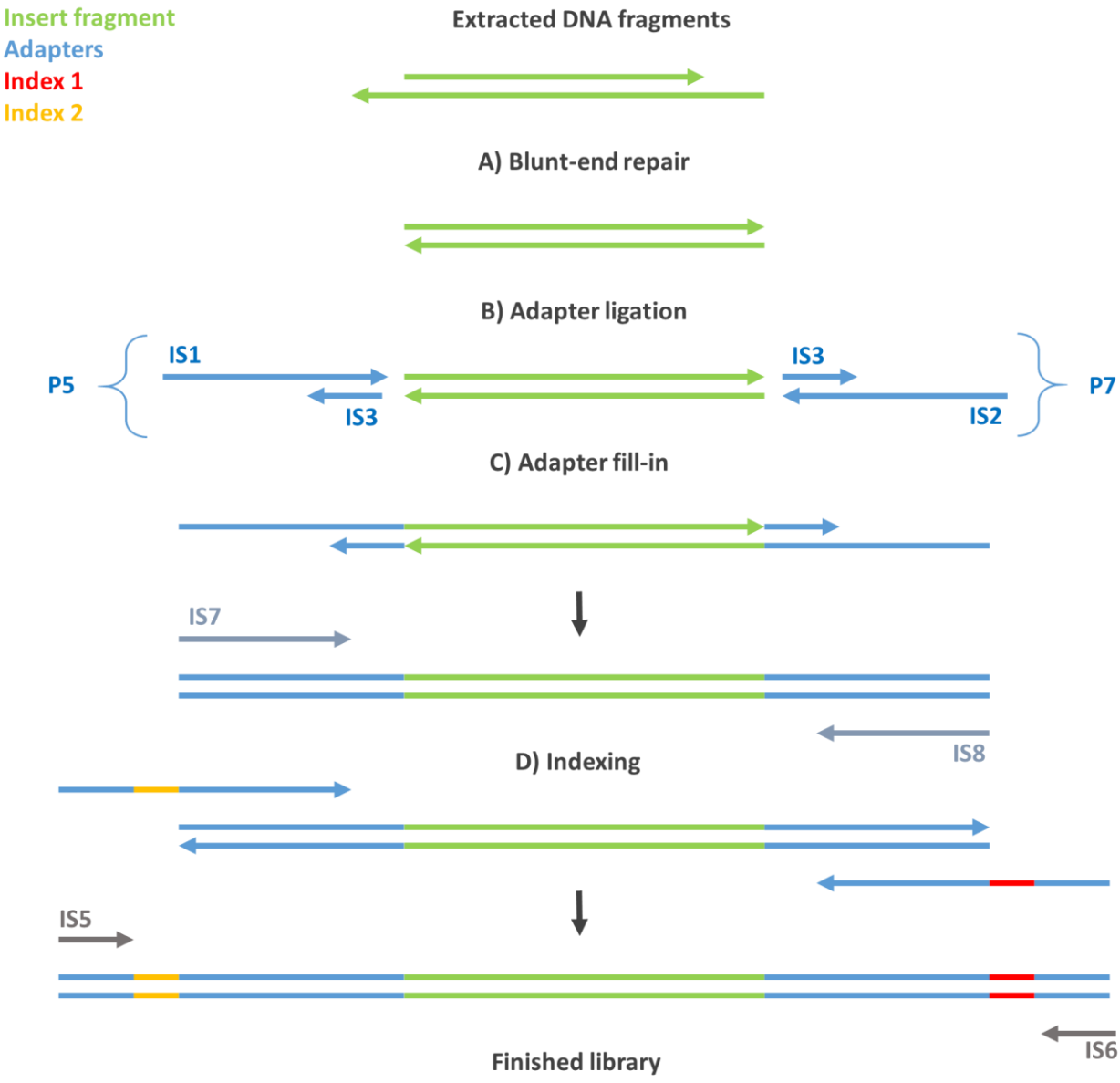

Figure 3: Overview of the library preparation protocol. Double-stranded fragments extracted from a sample are repaired to form blunt-end molecules. Adapters are then ligated to the 5'-ends of the fragments. Gaps in the 3' ends are filled in and full-length adapters are synthesised in the fill-in reaction. The resulting library can be amplified using IS7 and IS8 primers. A pair of unique indices is introduced in a PCR amplification with indexed primers, and the indexed library can be amplified using IS5 and IS6 primers. (Figure constructed after Kircher et al. (2012) and Meyer and Kircher (2010)).

First, single-stranded overhangs of the extracted DNA sequences were repaired to form blunt ends that allow subsequent adapter ligation (Figure 3A). In the blunt-end repair reaction, T4 DNA polymerase removes single-stranded 3' overhangs and fills in 5' overhangs by synthesising new DNA in the 5'–3' direction. T4 polynucleotide kinase (T4 PNK) transfers phosphate from ATP to the 5'-end of the DNA to make the blunt-ends available for subsequent adapter ligation. Master mix with 1X NEBuffer2, 100 µM dNTPs (Thermo Scientific), 0.8 mg/mL bovine serum albumin, 1 mM ATP, 0.4 U/µL T4 PNK (BioConcept AG), and 0.024 U/µL T4 DNA polymerase (BioConcept AG)

was prepared (Table A1.1). For each sample and negative extraction control, 40 μL of blunt-end repair master mix and 10 μL of DNA extract was mixed in a 0.2-mL PCR tube. For the negative library control, only master mix was added. Blunt-end repair was then induced by incubating the reactions in a thermal cycler at 15 °C for 15 minutes and then at 25 °C for 15 minutes.

After the blunt-end repair, the libraries were purified using MinElute PCR Purification Kit (Qiagen) to remove excess enzymes, buffers and other reagents. The whole reaction volume was transferred from the PCR tubes into MinElute spin columns and mixed with 200 μL of PB binding buffer (Qiagen). The columns were centrifuged at 14 000 rpm for 30 seconds in a tabletop centrifuge. The columns were washed twice using 600 of μL PE wash buffer and centrifuged as above. The columns were dry spun twice for one minute to remove all excess buffer. The columns were then transferred into new 1.5-mL Eppendorf tubes, and DNA was eluted with 18 μL of TET (TE with 0.025% Tween-20). TET was carefully pipetted onto the silica membrane, the columns were incubated for one minute at room temperature, and the DNA was eluted by centrifugation at 14 000 rpm for one minute.

Next, adapters P5 and P7 were added to 5′ ends of the blunt-end repaired DNA fragments (Figure 3B). Master mix with 250 nM adapter concentration was prepared by mixing 400 μL of 2X Quick ligase buffer (BioConcept AG) and 20 μL of premade 10 μM Solexa adapter mix. Solexa adapter mix contained 10 μM of each adapter (Sigma custom oligos). To each blunt-end repaired sample and negative control, 21 μL of adapter master mix and 1 μL of Quick ligase was added. Adapter ligation reactions were incubated at room temperature for 20 minutes. After that, samples were purified using MinElute purification protocol as described above, except using 20 μL of TET to elute the purified product.

The gaps between the insert fragment and the 3′ adapter were repaired using Bst polymerase (Figure 3C). For fill-in, master mix with final concentrations of 1X Isothermopol buffer (BioConcept AG), 1.25 nM dNTPs, and 0.4 U/μL Bst polymerase (BioConcept AG) was prepared (Table A1.2). For each sample, 20 μL of master mix was mixed with 20 μL of sample. Fill-in reactions were incubated at 37 °C for 20 minutes and then at 80 °C for 20 minutes. After that, 3 μL from each library was used to make 1:1 and 1:100 dilutions for the first qPCR quantification. Dilutions were transferred into the post-PCR laboratory, while the remaining 37 μL of the libraries were left in the clean room.

To assess library conversion efficiency, the libraries were quantified with a LightCycler 96 real-time quantitative PCR (Roche Life Sciences) using SYBR Green fluorescent dye (Thermo scientific)

and primers IS7 and IS8 (Sigma custom oligos), which amplify unindexed library molecules (Figure 3). The quantification relied on a standard series with known template quantities. A standard curve was calculated from the linear regression between the standard template quantity and the qPCR cycle number at which the SYBR Green signal from the standard was detected (Cq). Master mix for qPCR reaction was prepared by mixing 560 μL of 2X SybrGreen, 392 μL of water, and 56 μL of each primer, 10 μM IS7 and IS8 (Table A2.2). Standard series were diluted from the original standard mix that had 10^8 copies/μL, to make dilutions of 10^7, 10^6, 10^5, 10^4 and 10^3 copies/μL.

One μL of each library dilution and two replicates of each standard were mixed with 19 μL of qPCR master mix and added onto a 96-well optic plate. The plate was placed in the LightCycler 96 quantitative PCR machine and run for 40 cycles on following program:

Initiation:	95°C -- 10 min
Denaturation:	**95°C -- 30 sec**
Annealing:	**60°C -- 30 sec**
Extension:	**72°C -- 30 sec**
Melt Curve:	**60°C -> 95°C**

The DNA quantity in each library was extrapolated from the normalized mean quantities of the two dilutions.

Next, indices were introduced to the libraries in the clean room using 5'-tailed indexing primers (Sigma custom oligos; Table A2.1; Figure 3D). Two individually unique eight-base indices were added to each library by amplification with a hot start Pfu Turbo Polymerase (Agilent Technologies). Each library was split into four reactions to avoid amplification bias, making a total of 76 indexing reactions. Master mix containing free dNTPs and the Pfu Turbo Polymerase (Agilent Technologies) was prepared, adding the polymerase right before mixing the master mix with the samples (Table A1.3). Twenty μL of both indexing primers, 9.25 μL of the sample, and 86.7 μL of the master mix were mixed, and the indexing reaction was finalised in a thermocycler in the modern lab using the following program:

Initiation:	98°C -- 12 min
Denaturation:	**98°C -- 30 sec**
Annealing:	**58°C -- 30 sec**
Extension:	**72°C -- 60 sec**
Final extension:	72°C -- 10 sec

After indexing, the samples were again purified with MinElute purification described above. The first step (PB binding) was repeated four times using 500 μl of PB (and 100 μL of sample) to bind all DNA from four replicate indexing reaction into one column. The DNA was eluted with 50 μL of TE.

To assess final library conversion efficiency and to calculate the needed number of reamplification cycles, the libraries were again quantified with qPCR. Dilutions of 1:100 and 1:1000 were prepared from each library. The quantification was performed as before, except using primers IS5 and IS6 (Agilent Technologies) that only amplify molecules carrying indexed adapters (Table A2.2).

The indexed libraries were reamplified to reach a copy number of ~ 10^{13} or higher. All libraries, including the ones that already had a sufficiently high copy number, were reamplified to keep the workflow consistent. The required number of amplification cycles to achieve the desired copy number was calculated for each sample and they were divided into three groups. Samples in the same group went through the same number of amplification cycles.

A cold start Herculase II polymerase (Agilent) was used for the reamplification. Master mix with 1X Herculase II Reaction buffer (Agilent), 0.25 mM of each four dNTPs, 0.4 μM of IS5 and IS6 primers, and the Herculase II Fusion DNA Polymerase was prepared (Table A1.4). Herculase reactions were prepared on ice to keep the enzyme inactive.

To make one series of samples for shotgun sequencing and one for the mitochondrial enrichment, each sample was split into four twice, making eight replicate reactions from each library. Five μL of sample and 95 μL of Herculase master mix was aliquoted for each reaction into 200-μL PCR Eppendorf tubes. Reactions were divided into three thermocyclers corresponding with their assigned cycle number group, and the PCR amplification was run for the appropriate number of cycles (9, 13 or 18) with the following program:

Initiation:	95°C -- 2 min
Denaturation:	**95°C -- 30 sec**
Annealing:	**60°C -- 30 sec**
Extension:	**72°C – 30 sec**
Final Extension:	72°C -- 5 min.

The reamplified libraries were purified as before, using one column per four replicate reactions, two columns per sample. The purified products from both columns were eluted into one 1.5 ml Eppendorf tube with 10 μL of TET, yielding a total of 20 μL of eluate.

Next, DNA concentration of the reamplified and purified libraries were quantified using Agilent 2200 TapeStation System with High Sensitivity D1000 ScreenTape. To obtain a sequencing pool with equimolar ratios, 10 nM dilutions were made from each library and the dilutions were pooled together for shotgun sequencing.

3.5 Human mitochondrial enrichment

Due to the amount of endogenous DNA in ancient samples, target enrichment approaches are commonly utilised in aDNA studies to enrich the proportion of endogenous molecules (Carpenter et al., 2013). In hybridization capture, target DNA is captured with oligonucleotide probes that are designed to hybridize with the genomic regions of interest. I used in-solution hybridization capture with biotinylated single-stranded DNA probes following the protocol described in Furtwängler et al., 2018 (Figure 4).

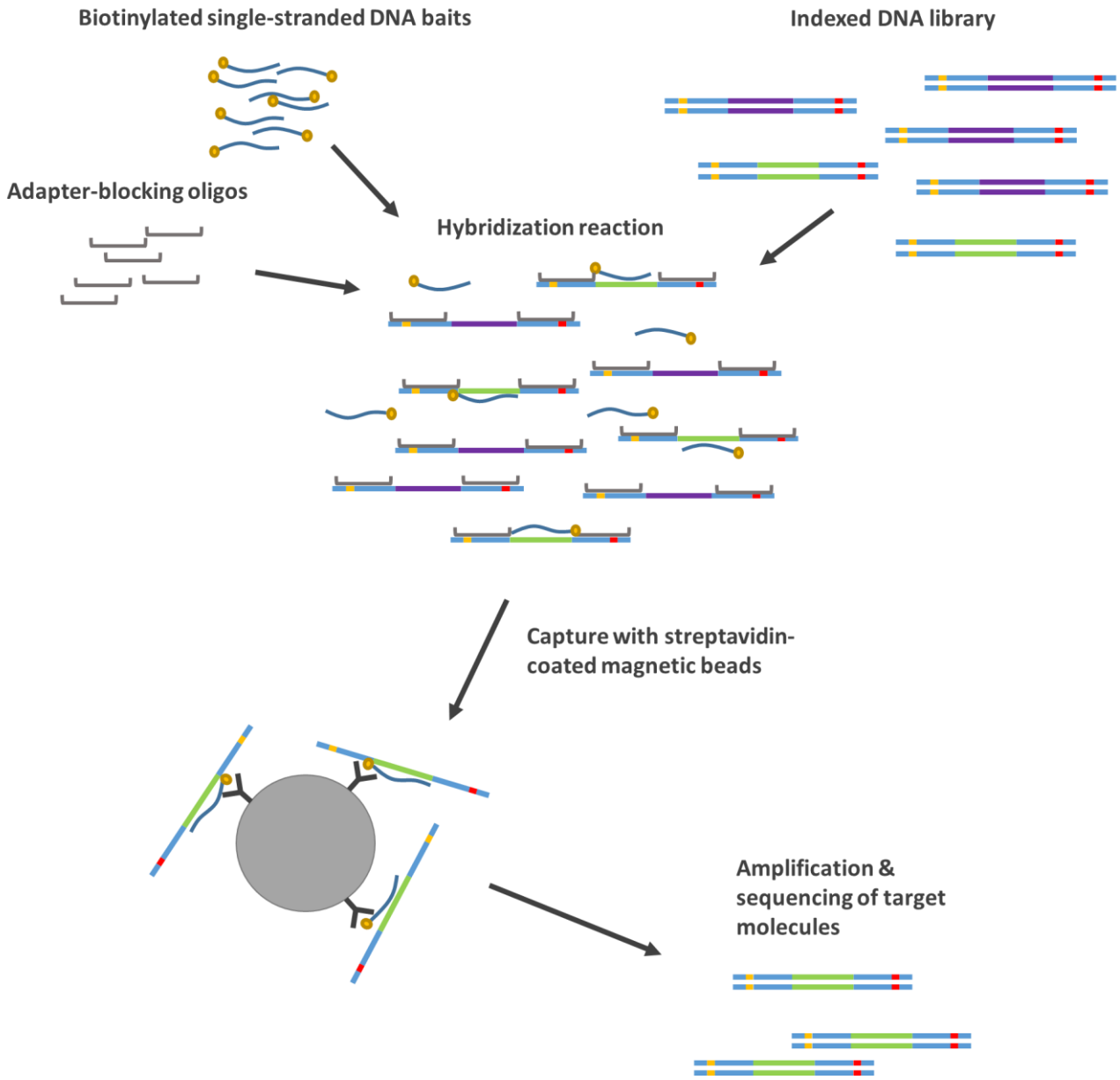

Figure 4: Target capture schematic. Pooled DNA libraries are hybridized with biotinylated single-stranded DNA probes specific for regions in target DNA (green). Small oligonucleotides are used to prohibit adapters from hybridizing with each other. Hybridized library molecules are captured with streptavidin-coated magnetic beads, and unhybridized DNA is washed away on a magnetic rack. Target molecules are then amplified on beads, purified and sequenced (Figure after Carpenter et al. (2013)).

Hybridization reactions were multiplexed by pooling DNA libraries into three hybridization pools. Only sample libraries were pooled and enriched here: negative controls were later pooled with control libraries from other experiments and captured and sequenced separately. The probe set used in the experiment targets genomic regions that cover the complete human mitochondrial genome and had been made in advance according to Maricic et al., 2010 and Fu et al., 2013 by IEM staff. Although the capture probes target human mitochondria, they would also likely enrich mitochondrial DNA of other mammals, albeit less efficiently (V. Slon, personal discussion). Obviously, a better solution would be to use a capture array that simultaneously targets several species, but that was not available for this study.

Before setting up the capture, all libraries had to go through an additional reamplification step, since the enrichment protocol required 2,000 ng of DNA per enrichment pool, and most of the samples were not concentrated enough. Ideally, we hoped to see concentrations of approximately 200 ng/μL or higher for each library. Most of the libraries had a significantly lower concentration, so a second Herculase amplification was carried out as above. The samples were again divided into eight separate reactions and the PCR program was run for 8 cycles.

After the second reamplification, the samples were again purified. To further concentrate the samples, we attempted eluting the DNA with a very small eluate volume: we eluted each of the two columns twice with 4 μL of TET (total of 16 μL eluate). However, subsequent TapeStation quantification showed that the concentration in all samples was lower than after the first reamplification. We suspected that the elution volume was too low and had not recovered all DNA bound on silica columns, and we thus eluted the same columns again with 10 μL of TET (total of 20 μL eluate), which is the lowest elution volume recommended by the purification kit manufacturer. We measured a subset of samples with TapeStation to see if the second elution would have sufficient concentration: it did not, but it was clear that we had not recovered all DNA with the first elution.

To further concentrate the samples, we did yet another Herculase amplification, this time splitting the libraries into four parallel reactions and using 4 cycles. Samples were purified as before, eluted into 10 μL of TET and quantified with TapeStation. The concentrations were still much lower than expected, and overall lower than they had been after the first Herculase amplification. We suspected that the amplification was too inefficient to compensate for the DNA loss in the purification and we decided not to compromise the samples with any further amplifications and prepared the enrichment pools from the libraries at hand. The samples were divided into three pools: the first

pool had six samples concentrated enough to reach the required total of 2,000 ng of DNA and equimolar sample ratios by increasing the reaction volume. The second and third pools (five samples each) were constructed from the poorer libraries simply by combining the remaining library volumes (9 µL each) (Table 3).

Table 3: Sample libraries divided into capture pools, DNA concentration of each library before pooling, and the added library volume.

	Sample	concentration (ng/µL)	µL to capture pool
POOL 1 Total 2,000 ng of DNA, equimolar sample ratios	ZH0724	219	1.52
	ZH0728	60.5	5.51
	ZH0729	80.9	4.12
	ZH0736	54	6.17
	ZH0738	56.1	5.94
	ZH0739	63.5	5.25
POOL 2 DNA quantitity NA, non-equimolar sample ratios	ZH0725	42.7	9
	ZH0730	44.9	9
	ZH0735	29.8	9
	ZH0732	NA	9
	ZH0737	NA	9
POOL 3 Total 679 ng of DNA, non-equimolar sample ratios	ZH0726	18.2	9
	ZH0727	6.26	9
	ZH0731	20.8	9
	ZH0733	9.21	9
	ZH0734	20.9	9

To prohibit the adapters from binding to each other during the capture, the adapters were masked using four adapter-blocking oligonucleotides (Sigma custom oligos; Figure 4). A mixture of blocking oligos was prepared by mixing 5.86 µL of each 500 µM blocking oligos, BO04, BO06, BO08, and BO10. Two µL of blocking oligo mixture was added for every 10 µL of sample library: 5.4 µL for the first pool and 9 µL for the second and third pools. Blocking oligos were hybridized with the libraries by incubating the reactions in a thermal cycler for 5 minutes at 95 degrees, 5 minutes at 65 degrees and 10 minutes at 35 degrees.

The hybridization capture buffer was prepared by combining 170 µL of preheated 60 °C 2X Hi-RPM hybridization buffer and 34 µL of 10X Agilent blocking agent (Agilent). Premade probes were pooled together by mixing 20 µL of each (mt1 and mt2), and the mixture was aliquoted into three empty 200-µL PCR tubes: 10 µL for the first pool and 15 µL for the second and the third pools. The hybridization buffer was then added on the probes: 50 µL for the first and 84 µL for the second and the third. Finally, library pools were added to their corresponding reactions, and the libraries were hybridized with the probes by incubating them in a thermal cycler for 48 hours at 65 °C.

For post-hybridization washes, 2xBWT (2M NaCl (Sigma-Aldrich), 10 mM Tris-HCl (pH 8; AppliChem), 1 mM EDTA (pH 8), and 0.1% of Tween-20) and HWT (1X PCR Gold Buffer (Applied Biosystem), 2.5 mM MgCl (Applied Biosystem), and 0.1% of Tween-20) wash buffers were prepared. For each pool, 20 μL of magnetic M-270 streptavidin beads (Life Technologies) were washed twice with 1000 μL of 1xBWT. The beads were immobilised on the tube wall with a magnetic rack during the washes, which enables the removal of the supernatant without disturbing the beads. The beads were resuspended to 20 μL of 1xBWT and 160 μL of 1xBWT and the hybridization reactions were then added onto the beads. Probes were captured on the beads by incubating them at room temperature in a rotator for 30 minutes. To remove unhybridized DNA, the beads were washed three times with 200 μL of 1xBWT and twice with 200 μL of 60 °C HWT. The HWT suspensions were incubated at 60 °C for 2 minutes. After that, the pools were washed once more with 200 μL of 1xBWT, transferred into new tubes, and washed with 100 μL of TET. Finally, the beads were resuspended into 15 μL of TET and transferred into new 200 μL PCR tubes.

Enriched library pools were quantified with qPCR and amplified on the beads in three parallel reactions with Herculase II. The amplified products were purified with MinElute columns as before, quantified with TapeStation, and pooled together in equimolar ratios for sequencing.

3.6 Sequencing

Both original ("shotgun") and enriched version of the sample libraries were paired-end sequenced with the Illumina HiSeq 4000 (2x75+8+8 bp) platform, which sequences 75 base pairs from both ends of the read and the 8 additional index nucleotides from both adapters. Sequences with unexpected index combinations were removed. Negative control libraries were also shotgun sequenced. Control capture libraries, which were prepared separately, were sequenced on a different sequencing run later, and thus could not be included in the analyses.

3.7 Validation of the bioinformatic analysis workflow

The bioinformatic analysis of the sequencing data was carried out following a modified version of the mammalian mitochondrial screening workflow introduced by Slon et al., 2017. The approach uses taxonomic classification algorithm for metagenomic data to assign each sequence a taxonomic label, which increases confidence of species detection from environmental samples and allows the screening of multiple species simultaneously. A major difference between the original version and the version presented here is the choice of software: here, I used Kraken for taxonomic labelling,

while Slon et al. used Megan (Huson et al., 2007; Wood and Salzberg, 2014). The main advantage of Kraken over Megan is its speed. Unlike Megan, which requires Blast input, Kraken is a standalone program. It also enables easy building of custom databases.

I first assessed the performance of the two programs using simulated datasets, since Kraken had not been benchmarked with ancient data. I performed the assessment as an internship project at the Max Planck institute for Evolutionary Anthropology, Leipzig, Germany, during March 18–April 17, 2019. The project's primary goal was to evaluate if Kraken could replace Megan in the institute's local analysis pipeline for cave sediments. Relevant parts of the work are presented here.

Kraken is a taxonomic classifier for short DNA sequences (Wood and Salzberg, 2014). It relies on exact matching of short subsequences, *k*-mers, instead of inexact alignment of a full sequence, such as Blast-like algorithms do. A database of *k*-mers, where *k* is the length of the subsequences, is built from a set of reference sequences. The user can define both the *k*-mer length and the reference sequences that are included in the database. To assign a taxonomic label for a sequence, each *k*-mer within a query sequence is searched against the database, and a *k*-mer is assigned to its lowest common ancestor (LCA) in the NCBI taxonomy, i.e. to the node that is ancestral to all taxa that share the given *k*-mer (Figure 5). A pruned subtree of all classified *k*-mers within a query sequence is then used to score all possible root-to-leaf-paths (RTL) to give the sequence a taxonomic label.

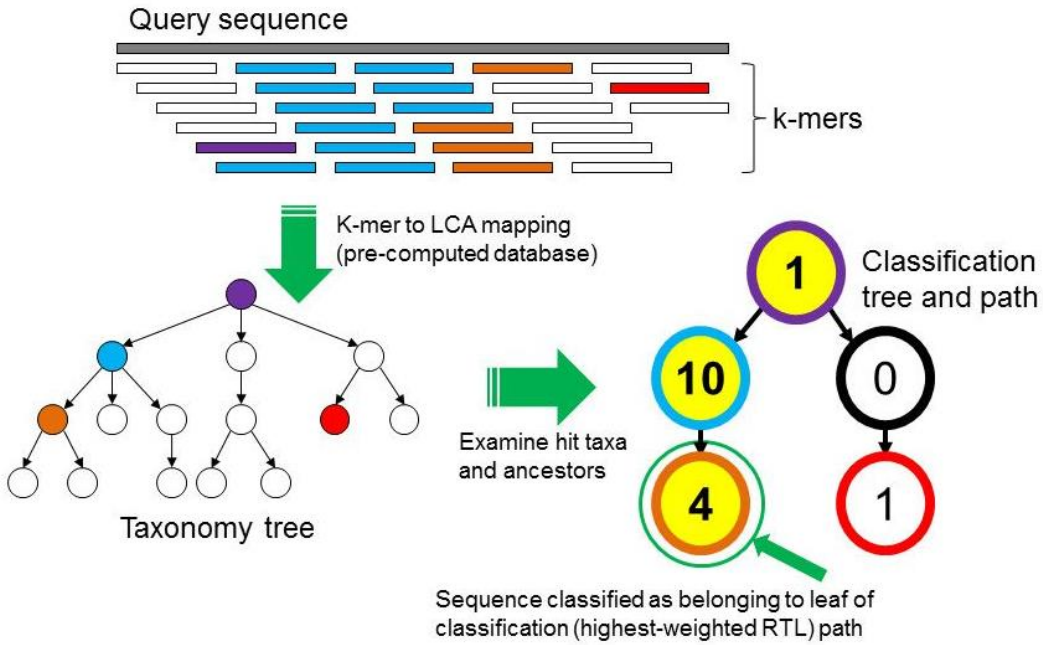

Figure 5: Schematic of Kraken's classification algorithm (adopted from Wood and Salzberg, 2014 *under CC-BY 4.0 license). First, Kraken maps all* k-mers *within a query sequence to their LCA in the taxonomic tree. A pruned subtree, which contains the nodes associated to the query sequence's* k-mers *and their ancestors, is then used to weight all RTL paths within the subtree. A taxonomic label is assigned based on the highest scoring RTL path. If more than one RTL paths have equally high scores, the query sequence is assigned to their LCA.*

I assessed Kraken's performance using three simulated datasets generated by Slon et al. (2017), which contained either 100%, 1% or 0% mammalian mitochondrial sequences (Table 4). The dataset consists of short sequences generated from five published mammalian mtDNA genomes (from spotted hyaena, cattle, pig, mammoth and Neandertal) and 114 bacterial genomes. Because the reference database used in taxonomic labelling consists of NCBI reference genomes, the simulated data was constructed from non-reference mammalian genomes to mimic divergence. The sequences had 50% terminal C-to-T substitutions on either none of the taxa (simulation A), all taxa (simulations B and E), or a subset of the taxa (simulations C and D), corresponding with the damage frequencies usually observed in Pleistocene aDNA samples. Additionally, simulation E carried 10% of C-to-T substitutions in other than terminal positions. Each simulation consisted of 100,000 sequences and had a uniform length distribution between 35 and 100 nucleotides.

Table 4: Description of the simulated dataset.

Family	**Bacteria**	**Hyaenidae**	**Bovidae**	**Suidae**	**Elephantidae**	**Hominidae**	**Types**
Species	114 species	Crocuta crocuta	Bos taurus	Sus scrofa domesticus	Mammuthus primigenius	Homo sapiens neanderthalensis	
GenBank ID	NA	JF894377	DQ124371	KC469586	EU153448	KC879692	
Mammals only	0%	50%	30%	15%	4%	1%	A, C, E
1% mammals	99%	0.50%	0.30%	0.15%	0.04%	0.01%	A, B, C, D, E
Bacteria Only	100%	0%	0%	0%	0%	0%	A, B, E

Simulation damage types	
A:	No damage.
B:	50% terminal C-to-T changes in all genomes.
C:	50% terminal C-to-T changes in Hominidae, Hyaenidae and Elephantidae.
D:	50% terminal C-to-T changes in bacteria, Hominidae, Hyaenidae and Elephantidae.
E:	50% terminal C-to-T changes in all genomes + 10% chance of C-to-T change in any other position.

By default, Kraken uses *k*-mer length 31, but that was considered too long for ancient DNA sequences. To find an optimal *k*-mer length for the simulated data, I built five custom Kraken databases with varying *k*-mer lengths. the databases consisted of the same set of mammalian reference mitochondrial genomes that Slon et al. (2017) used to test Megan: the reference set has a comprehensive array of 791 NCBI reference mitochondria genomes across mammalian taxa (Appendix 3). The databases were built for *k*-mer lengths 16, 20, 24, 28 and 30, and the taxonomic classification of the simulated sequences was repeated with each database version.

The proportion of reads assigned to expected (Hyaenidae, Bovidae, Suidae, Elephantidae and Hominidae) and unexpected (other) families was recorded and compared to the outcome of Megan. Furthermore, I estimated Kraken's accuracy using the Neanderthal sequences as a representative subset of the simulated datasets. I calculated sensitivity and precision on a family level (Hominidae) and compared the outcomes of Kraken and Megan. Here, I define sensitivity as a proportion of correctly classified Neanderthal sequences out of all Neanderthal sequences in the data. Precision is defined as the proportion of correctly classified Neanderthal sequences out of all sequences assigned to Hominidae. Based on the results, I chose to use the database built with *k*-mer length 24 to analyse the real data

3.8 Analysis of the sequencing data

The sequences retrieved from the Finnish sediment samples were analysed utilising Kraken and the mammalian mitochondrial database built with *k*-mer 24. First, the sequenced reads were prealigned to a small set of mammalian mitochondrial reference genomes to remove the majority of environmental sequences and to make taxonomic classification faster (Figure 6). Mapped reads were then classified with Kraken. Classified reads were split by family and sequences assigned to each family were realigned to one (or more) relevant reference. Finally, the authenticity of the sequences mapped to each reference was evaluated to decide whether the detected taxa is ancient or not.

The analysis steps before and after taxonomic classification were carried out using the EAGER (Efficient Ancient Genome Reconstruction) pipeline (Peltzer et al., 2016). The pipeline is designed for ancient DNA analysis, and it implements both ancient DNA specific tools and standard next generation sequence analysis software. All of the steps from sequence preprocessing through quality control, mapping and variant calling could be carried out with EAGER. Here I used EAGER version 1.9 for quality control (FastQC), adapter removal and read merging (AdapterRemoval v2), mapping (BWA and CircularMapper), deduplication (DeDup), and post-mortem DNA damage calculation (mapDamage) (Figure 6).

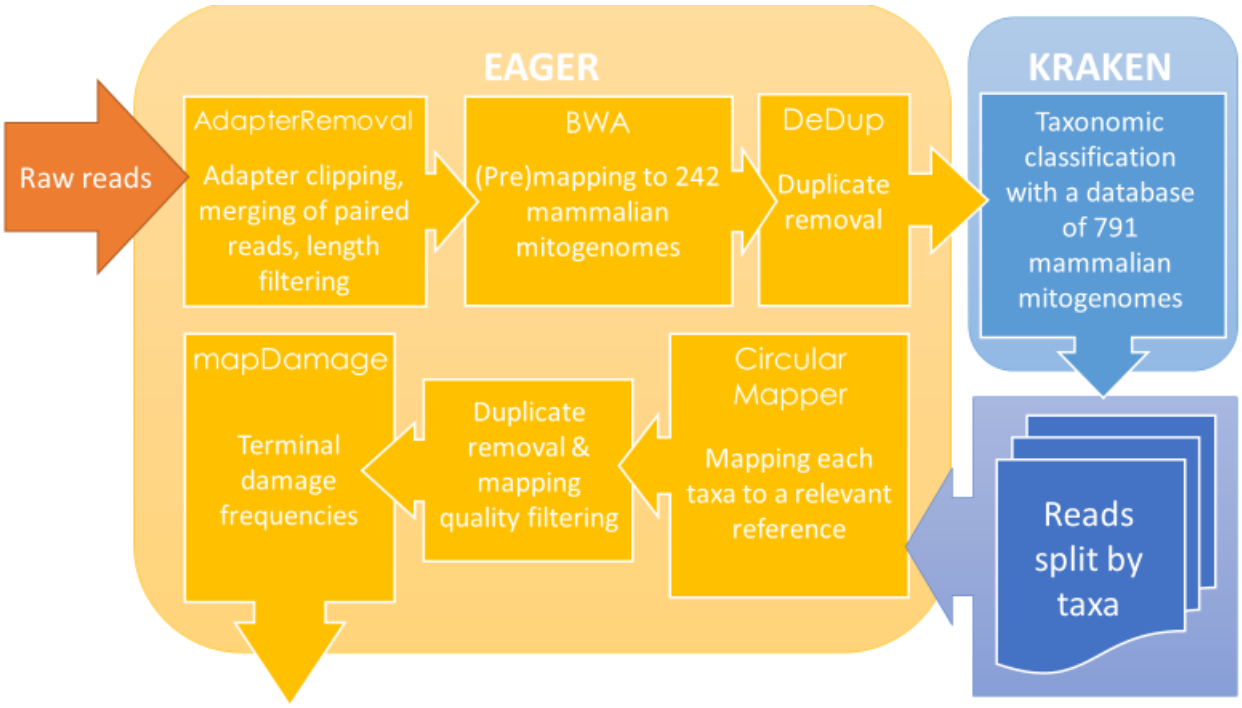

Figure 6: Outline of the analysis workflow used here. Within EAGER, sequences are reprocessed with AdapterRemoval, mapped against reference array with BWA and deduplicated with DeDup. Taxonomic classification is carried out separately with Kraken, and classified reads are split by taxa with an in-house script. Individual taxa are then mapped to a relevant reference genome, deduplicated, filtered and authenticated using EAGER.

Quality control

I used FastQC to assess the quality of the raw reads and preprocessed reads (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). FastQC analysis was run with default options for raw forward and reverse reads separately, and again for merged reads after adapter removal, merging, and length filtering. Results were summarised and visualised with MultiQC (Ewels et al., 2016).

Adapter removal and merging

In ancient DNA studies, sequenced reads are typically very short, because the extraction protocol targets short DNA fragments (Schubert et al., 2016). Consequently, adapters often become partially sequenced with the insert and must be removed before further analysis. When paired-end sequencing is used, another consequence of the short fragment length is that forward and reverse reads overlap. Subsequently, paired reads can be merged based on the overlapping region to reconstruct the original fragment, which is then aligned as a single read. Merging increases confidence of the base calls on the overlapping region, retains both ends of the fragments, and makes alignment easier.

Adapters were removed and paired sequences were merged using AdapterRemoval v2 (Schubert et al., 2016). By default, AdapterRemoval trims 5' terminal bases below base quality 20. I applied 35 bp as a sequence length cutoff. Very short reads are more likely to match the reference genome by chance (Smith et al., 1985) and are often excluded from the analysis to avoid spurious alignment of microbial sequences (de Filippo et al., 2018). A length cutoff of 30-35 bp is commonly used in ancient DNA studies (Dabney et al., 2013b; Meyer et al., 2016) Here, the read length cutoff was set to 35 bp, because Kraken had not been tested on reads shorter than that.

Mapping

To filter out most of the environmental sequences, trimmed and merged reads were first mapped against a representative set of 242 mitochondrial reference genomes listed by Slon et al. (2016). Three entries (NC_006923: Bradypus variegatus (Brown-throated sloth), NC_009574: Mammut Americanum (American mastodon) and NC_001807: Homo sapiens) had been removed from the NCBI database since the publication of the paper, and they were replaced with updated versions: NC_028501 for B. variegatus, NC_035800 for M. americanum and NC_012920 for H. sapiens. All records were downloaded from NCBI Nucleotide on 4.4.2019.

Reads were aligned using Burrows-Wheeler Aligner (BWA) with EAGER's default parameters, which are adjusted for ancient DNA by allowing more gaps than normal (Li and Durbin, 2010; Peltzer et al., 2016). For comparison, mapping was also performed with even more permissive parameters that may improve the mapping of ancient fragments with post-mortem damage: this was achieved by choosing a seed length significantly longer than the reads length of the data (Meyer et al., 2012). However, seed length adjustment did not improve mapping, but rather decreased the number of mapping reads approximately by half. Reads mapped with default parameters were therefore carried over to the downstream analyses.

Duplicate removal

Duplicates were removed with DeDup, EAGER's native deduplication tool (Peltzer et al., 2016). As opposed to e.g. Samtools rmdup, which takes into account only the start coordinates of mapped reads, DeDup also considers the end coordinates, which helps to distinct true duplicates when overlapping paired-end reads have been merged.

Taxonomic classification

Next, trimmed, merged and deduplicated reads mapped to any of the 242 reference genomes were classified with Kraken. The database of 791 NCBI reference mitochondrial genomes from various mammals, built using *k*-mer length 24 as described above, was used to assign each sequence a taxonomic label. All reads classified to family level or lower were extracted and split by family. I required a family to have at least 1% of all classified reads and a minimum of 10 reads for a given family to pass the detection threshold. Reads assigned to families below that threshold were excluded from further analysis.

Realignment and damage profiling

Reads assigned to families that passed the detection threshold were realigned to an individual relevant reference mitochondrial genome from the database. A suitable reference genome for each family was selected based on Kraken's species level assignments. When species level assignments were inconclusive, the geographically most reasonable candidate was chosen. For Hominin reads, I used the revised Cambridge Reference Sequence of the human mtDNA.

EAGER was again utilised for remapping, deduplication and terminal damage profiling. For mapping, I used EAGER's native tool, CircularMapper. It relies on BWA but creates an elongated version of a circular reference genome, which improves the mapping on the ends of the linearised reference (Peltzer et al., 2016). Mapping quality cutoff was set to Phred score 25. DeDup was run as described above. Finally, terminal damage frequencies were calculated and visualised with mapDamage (Peltzer et al., 2016).

The number of mapping reads, coverage distribution, terminal damage frequencies, and mean fragment length were used to evaluate ancient DNA authenticity. Reads were expected to map evenly across the reference, and approximately 100 reads were required to confidently calculate the damage frequencies. To call DNA ancient, at least 10% of terminal positions were expected to carry a C-to-T substitution, and the mean fragment length was expected to be below 70 bp.

4 RESULTS

4.1 Human mtDNA recovery and ancient DNA authenticity

To test whether ancient human DNA could persist in Finnish archaeological sediments, we extracted DNA from soil samples from Stone Age settlement sites and converted the extracts into double-stranded Illumina sequencing libraries, which we enriched for human mitochondrial DNA. Sequences were successfully obtained from all DNA libraries, and the sequencing data was analysed with an ancient DNA analysis pipeline and a taxonomic classifier. My main focus was on the sequences assigned to family Hominidae, which were realigned against the human mitochondrial reference genome.

Hominidae passed the detection threshold in five of the samples enriched for human mitochondrial DNA (Table 5). The number of reads assigned to Hominidae ranged from 11 to 333 and most of the assigned reads could be successfully realigned to the human mitochondrial reference genome. Two samples, ZH0725 from Spångkärret and ZH0736 from Kammarlahti, had over 100 mapping reads. In addition, another sample from Kammarlahti, ZH0738, had 58 mapping reads. Since all Kammarlahti samples came from one larger sample, it was reasonable to combine the data from the two samples before calculating damage frequencies and coverage distribution. We also detected human mtDNA in samples ZH0728 from Karpankangas and ZH0734 from Taipaleenranta, but the number of reads was well below 100 and thus their authenticity could not be estimated.

Table 5: Statistics for the samples passing the detection threshold for human DNA.

Site	Sample Name	# reads		Coverage		Terminal damage				Fragment length		GC%
		Assigned to Hominidae	Mapped, dedupped & Q25 filtered	Mean	std. dev.	1st Base 3'	2nd Base 3'	1st Base 5'	2nd Base 5'	Average	Median	
Spångkärret	**ZH0725**	**333**	**316**	**1.4**	**1.5**	**0.13**	**0.08**	**0.08**	**0.06**	**74**	**74**	**45.3**
Karpankangas	ZH0728	11	8	0.0	0.2	0.00	0.00	0.00	0.00	65	72	46.3
Taipaleenranta	ZH0734	22	18	0.1	0.3	0.00	0.00	0.00	0.00	73	71	43.2
Kammarlahti	ZH0736	300	233	1.4	2.3	0.00	0.00	0.00	0.00	97	96	44.6
Kammarlahti	ZH0738	69	58	0.3	1.0	0.00	0.00	0.00	0.00	88	84	44.9
Kammarlahti, combined	**ZH0736, ZH0738**	**369**	**291**	**1.7**	**2.5**	**0.00**	**0.00**	**0.00**	**0.00**	**95**	**95**	**44.7**

Coverage distribution across the human mtDNA reference was visualised for the Spångkärret sample and the combined samples from Kammarlahti using Integrative Genomics Viewer (Robinson et al., 2011) (Figure 7). While Spångkärret had relatively uniform distribution, Kammarlahti had clusters of reads with nearly identical start and end coordinates, which indicates high duplication levels and inefficient duplicate removal; indeed, enriched libraries from Kammarlahti showed an extremely high cluster factor (see 4.5 Data quality and contamination).

Kammarlahti showed no signs of terminal damage (Figure 8). In addition, the average fragment length of the human mtDNA sequences was 95 base pairs, clearly longer than expected from ancient DNA, further suggesting that the detected human DNA originates from modern contamination. The sample from Spångkärret, which had the highest number of mapping reads of all samples, showed evidence of terminal damage: 8% for the first base in the 5′ and 13% in the 3′ end. The frequencies are close to the minimum of 10% deamination frequency and indicate at least some level of DNA degradation. The average fragment length was 74 base pairs, which is shorter than in samples from Kammarlahti, but still longer than usually seen in authentic ancient DNA.

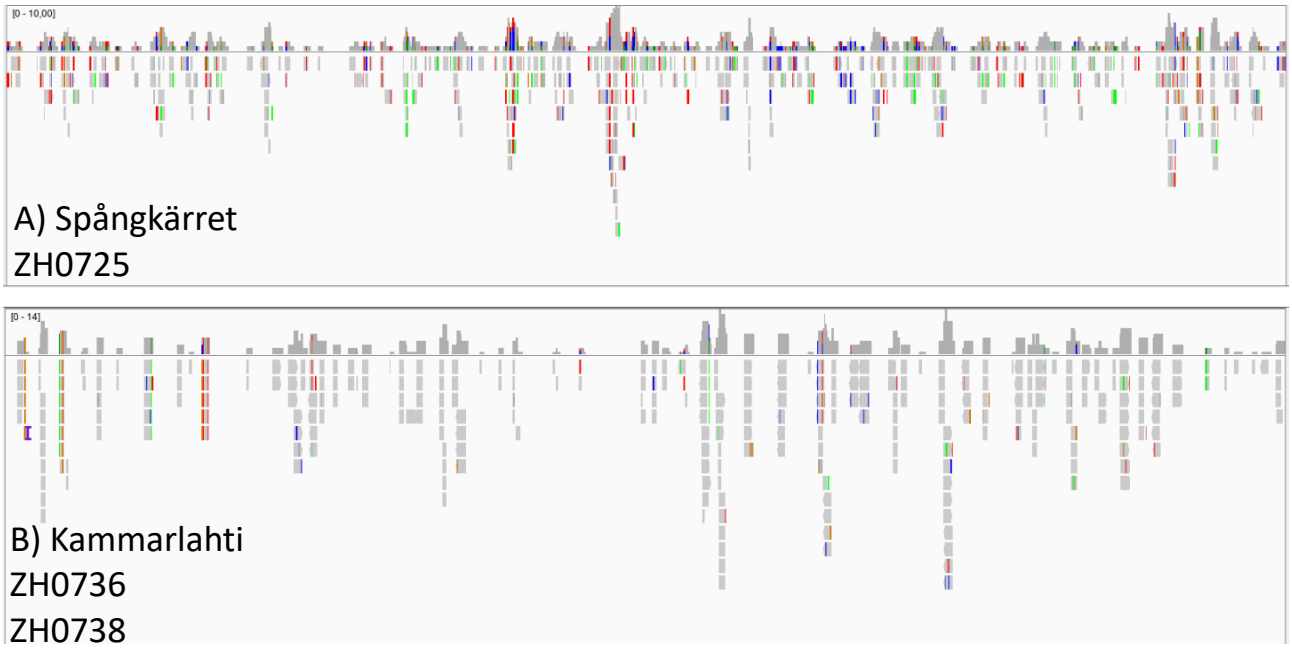

Figure 7: Coverage across the human mitochondrial reference genome in samples with > 100 mapping reads. A) A sample ZH0725 from Spångkärret shows relatively even distribution. B) The combined samples from Kammarlahti (ZH0736 and ZH0738) appear to have nearly identical reads covering the same positions several times.

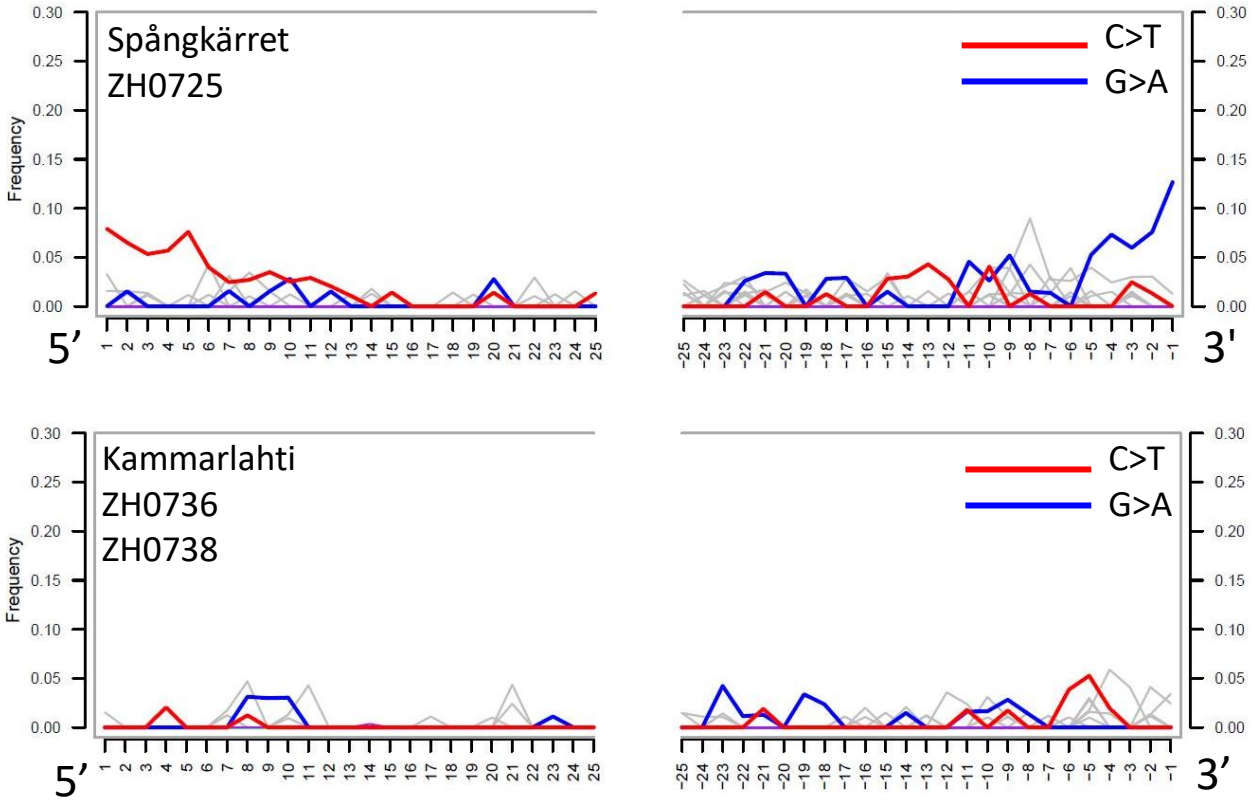

Figure 8: Damage patterns in the samples with over 100 reads assigned to human: Spångkärret and two combined samples from Kammarlahti. Spångkärret sample shows increase of C-to-T substitutions in the 5’ end and corresponding G-to-A substitutions in the 3’ end, indictive of DNA degradation. Kammarlahti has no signs of terminal damage.

No other mammalian species apart from human were detected in any of the samples. Originally, I detected Mustelidae in almost all of the samples, but it turned out to be a false positive: reads assigned consisted almost exclusively of AC-repeats, strongly suggesting that they did not really come from a mustelid. Nevertheless, I realigned the reads to the mitochondrial reference genome of European badger (*Meles meles*), which was the most common species level assignment within the family and geographically the most probable candidate. As expected, all reads were aligned to a single low complexity region near the end of the badger's mitochondrial genome, confirming that the detection of Mustelidae in the samples was caused by spurious assignment of the low-complexity sequences.

4.2 Kraken: results from simulated data

Overall, Kraken performed well with the simulated data. Simulated terminal substitutions had very little effect on the classification performance, while internal errors (simulation E) clearly had an adverse effect on the number of classifications and classification accuracy. Databases built with *k*-mer length 20 or longer all produced reliable results, while the database built with *k*-mer length 16

produced a high number of false positives: in simulations with only bacterial sequences, almost 10% of reads were incorrectly assigned to mammals on family level (Figure 9). The same pattern was observed in simulations with 99% bacteria (data not shown). However, the false positive rate decreased drastically between *k*-mer 16 and 20 and did not change significantly after *k*-mer 24. Therefore, *k*-mer 16 was deemed too short and removed from the further comparisons.

In simulations with 100% and 1% of mammalian sequences, no differences were seen in the classification outcomes between *k*-mers 24 and 30 when only terminal damage was considered (Figures 10 and 11). However, the simulation E with internal damage had less assignments with longer *k*-mers. Overall, Kraken classified more reads to expected families than Megan did, except with simulation E and the *k*-mer 30 database, where Megan outperformed Kraken. Although Kraken assigned more reads in general, Megan assigned less reads to families not present in the simulation data, suggesting higher accuracy for Megan.

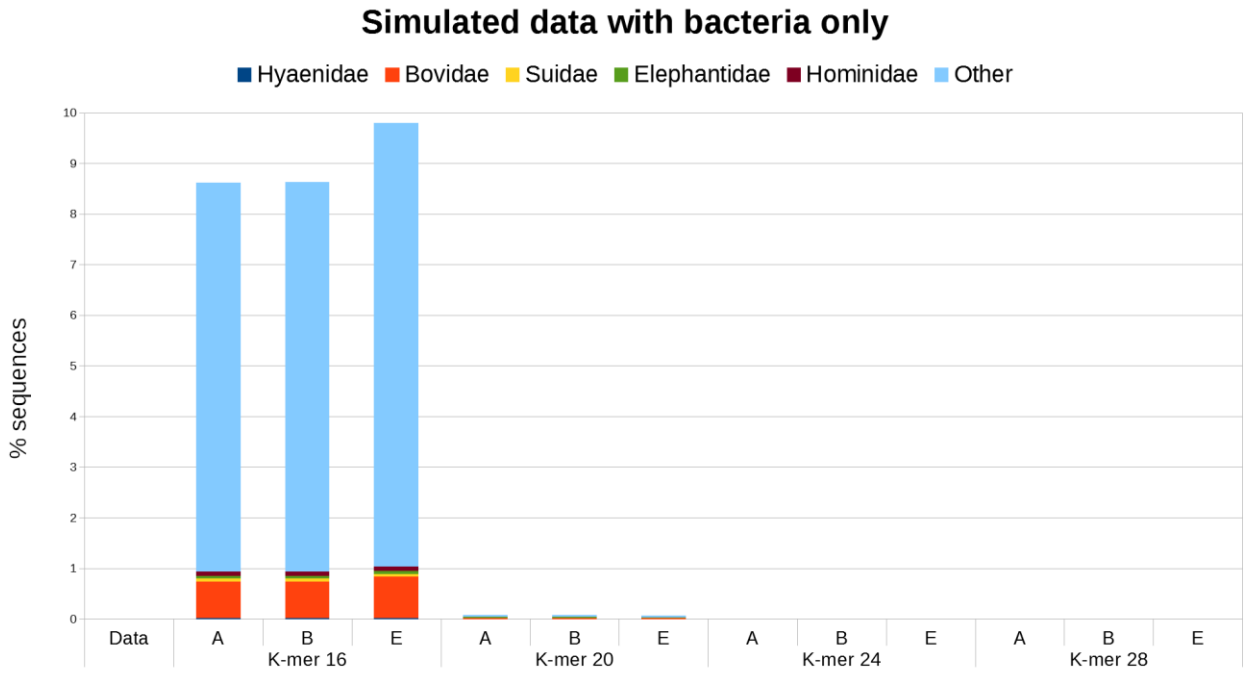

Figure 9: The proportion of reads assigned to mammalian families in simulated data consisting of only bacterial sequences. No classifications were expected. The database with k-mer length 16 produced high number of false positives. K-mers above 24 and MEGAN produced no false positives (k-mer 30 and MEGAN comparisons not shown).

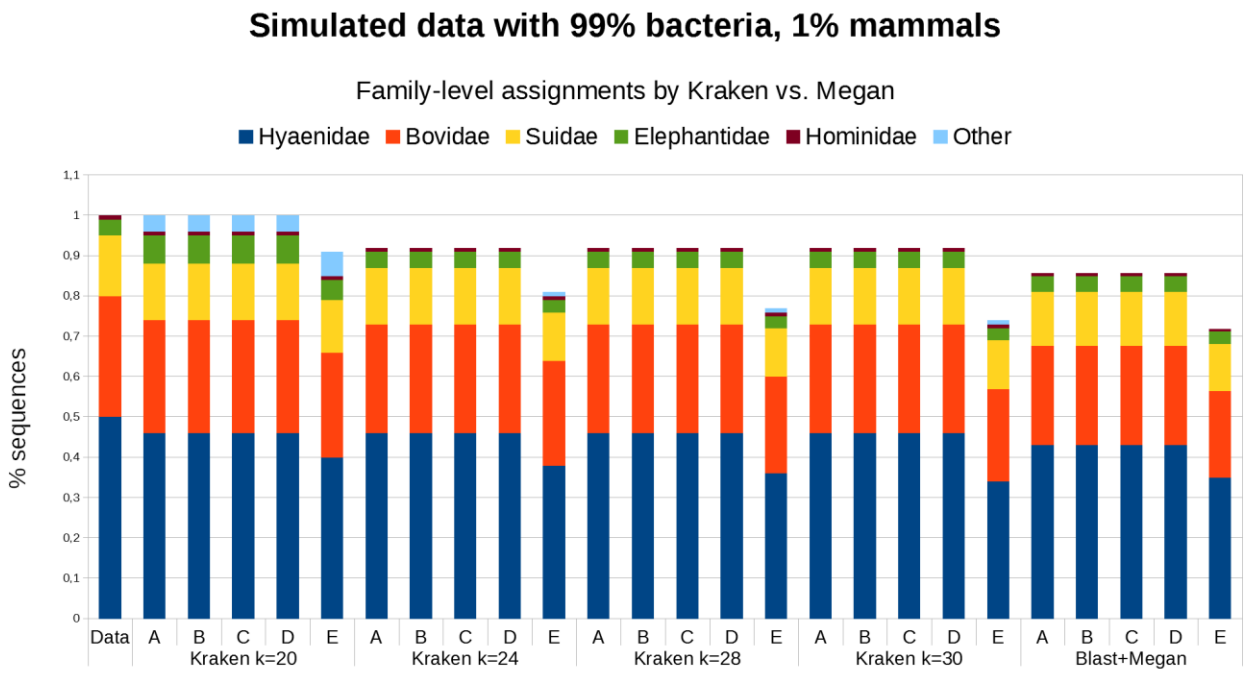

Figure 10: The proportion of reads assigned to expected and unexpected ("others") mammalian families by Kraken with k-mers 20-30 and MEGAN, in simulations with 1% mammalian sequences and 99% bacteria. Simulations A-D show no difference after k-mer 24. Simulation E has less classifications with longer k-mers. K-mer 20 still produced some false positives, represented by the presence of "other" families and the excess reads assigned to Elephantidae.

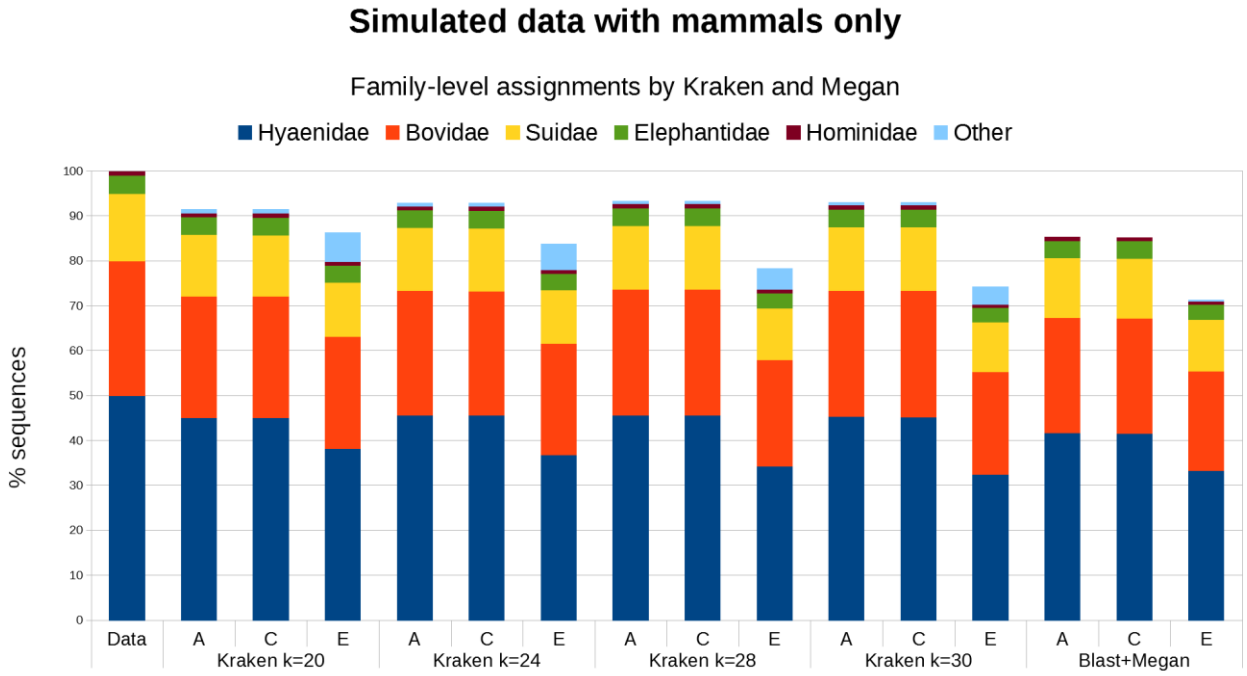

Figure 11: The proportion of reads assigned to expected and unexpected ("others") mammalian families by Kraken with k-mers 20-30 and MEGAN, with simulated data consisting solely of mammalian sequences.

Accuracy was further estimated using Neanderthal sequences from simulations with 100% mammalian sequences as a proxy. When the simulation E with internal damage was excluded,

Kraken was more sensitive than Megan and equally precise with all *k*-mer lengths (Figure 12). Both programs were less accurate when sequences had internal damage. With internal damage, Megan slightly excelled over Kraken in precision, but Kraken was more sensitive with medium *k*-mer lengths and the absolute number of correctly classified reads was thus higher with Kraken. Additionally, Kraken classified a small portion of hominin reads to incorrect families, which did not occur with Megan. Because *k*-mer 24 was the longest *k*-mer that correctly classified more reads than Megan with all simulations, it was chosen to be used in the analysis of the real data (Table 6).

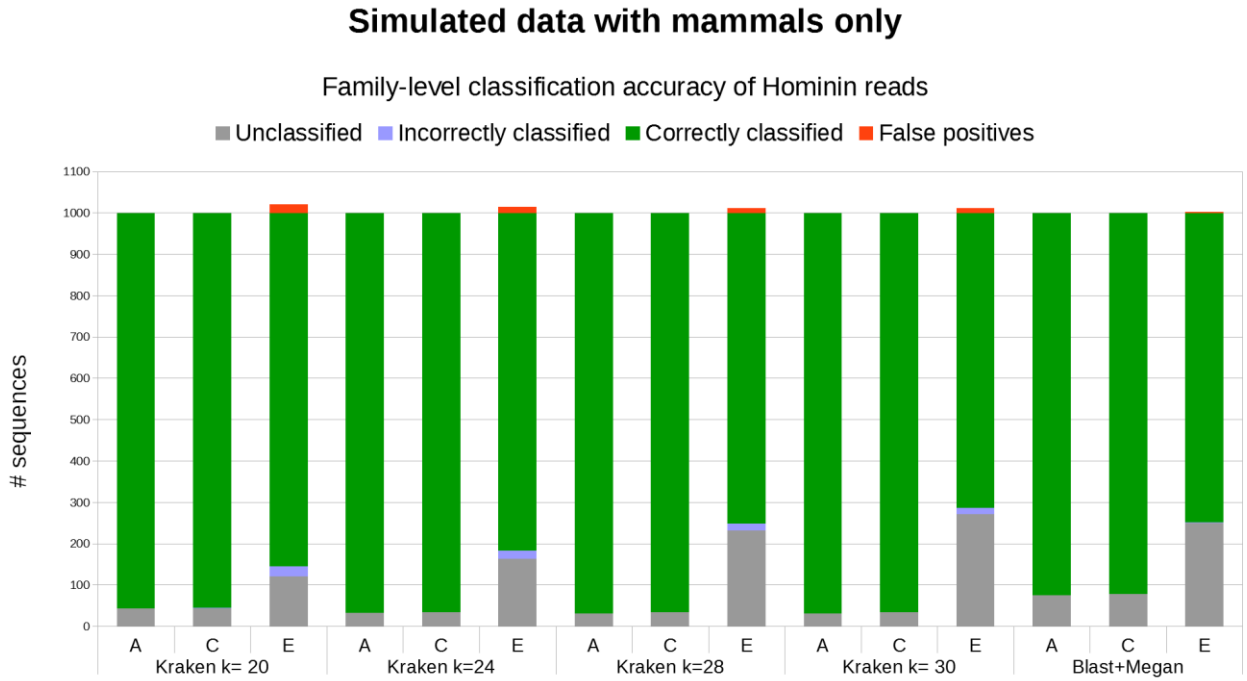

Figure 12: Kraken's classification accuracy compared with Megan on family level, with Neanderthal sequences and classifications on Hominidae family as a proxy. With no damage (A) or only terminal damage (C), Kraken classifies more read correctly. With internal damage (E) Megan is more precise, but Kraken with k-mers 20 and 24 is more sensitive. Kraken also classified a small number of Neanderthal sequences to incorrect families, which was not observed with Megan.

Table 6: Comparison of family-level accuracy in Kraken with k-mer 24 and Megan.

SIMULATION	NO DAMAGE (A)		TERMINAL DAMAGE (C)		TERMINAL + INTERNAL DAMAGE (E)	
PROGRAM	**Kraken**	**Megan**	**Kraken**	**Megan**	**Kraken**	**Megan**
FAMILY-LEVEL SENSITIVITY	**96.6%**	92.4%	**96.4%**	92.1%	**81.5%**	74.7%
FAMILY-LEVEL PRECISION	100.0%	100.0%	100.0%	100.0%	98.1%	**99.5%**

Although no false positives on Hominidae were seen in the simulations with no damage or terminal damage only, the simulation with internal damage always produced some false positives. Therefore, I checked if the false positives would persist through realignment, the next step in the analysis workflow. For that, I used the mammalian-only simulation data that had been classified with Kraken and the chosen *k*-mer 24 database. Sequences assigned to Hominidae were extracted

and realigned against the Neanderthal reference mitochondrial genome using BWA with default parameters. Remapping efficiently pruned out all false positives (Figure 13). In simulations with no damage (A) and only terminal damage (C), BWA was able to align nearly all sequences. However, it failed to map a substantial number of true positives when sequences carried internal damage (E).

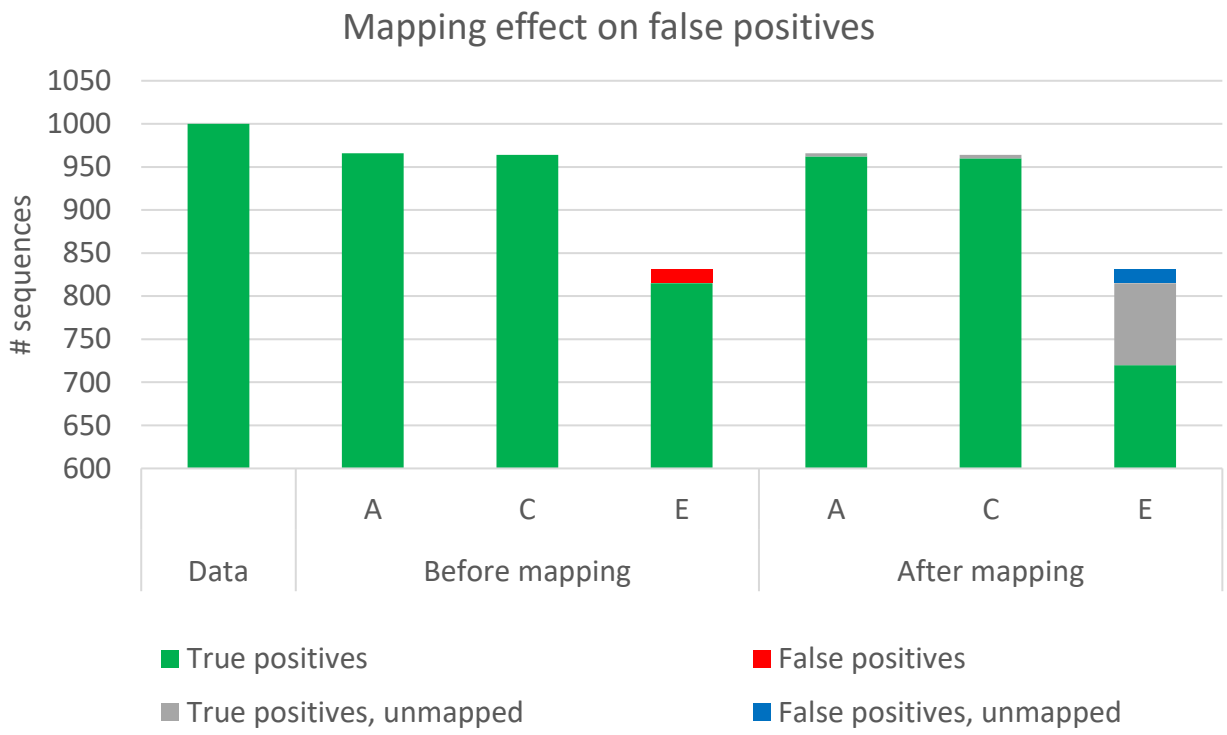

Figure 13: The fate of reads assigned to Hominidae after remapping to a relevant reference. Most true hominin reads with no damage or terminal damage only were successfully mapped to the reference, while almost half of true positives with internal damage were unmappable. On the other hand, no false positives with internal damage survived through remapping.

4.3 Inhibitor carry-over in DNA extraction

To extract DNA from the sediment samples, we applied a widely used and thoroughly tested ancient DNA extraction protocol suitable for various substrates. It is designed to efficiently remove inhibitory substances while preserving as many small DNA fragments as possible, but it is not optimised for soil. We noted that many of the samples produced substantially dark extracts after the digestion: the extract colour varied from almost clear to dark brown or black. While few sediment samples carried visible coal particles, which explained the dark colour of their extracts, some light-coloured sediments also produced surprisingly dark extracts. Darker colour could indicate the presence of particles, but also humic acids, which would strongly inhibit polymerase reaction. Excluding the samples that carried coal, there seemed to be some level of consistency in extract colour within sites: the coast sites, Spångkärret, Niskasuo, and Karpankangas, generally produced

darker extracts than Konstunkangas, Taipaleenranta, and Valkeakoski, while the samples from Kammarlahti were almost clear.

Despite purification, several of the darkest extracts carried some hint of colouration all the way to the library preparation, which could indicate a considerable carry-over of soil-derived substances in the extraction. Moreover, the libraries that produced darker extracts seemed to have lower copy numbers in the qPCR quantifications during the library preparation, but the pattern disappeared after Herculase amplification.

4.4 Library quantification and amplification efficiency

Overall, the DNA extracted from the sediment samples was converted into DNA libraries with moderate efficiency. Before indexing, the molecule copy number in the sample libraries ranged from ~ 10^5 and ~ 10^9 copies/µL (Appendix 4). The samples in the lower end of the copy number range indicated borderline acceptable library conversion efficiency or low DNA quantity in the extract. However, the mean was in the order of ~ 10^8 copies/µL, indicating a successful library preparation for at least the majority of the samples. After indexing, the copy number was determined to be between ~ 10^7 and ~ 10^{11}, with a mean of ~ 10^{10}, which indicated a moderate indexing efficiency in the best samples. Negative controls had a mean of ~ 10^6 copies/µL after indexing, indicating low rates of laboratory contamination.

Library quantification relies on linear regression calculated based on known template concentrations of qPCR standard series, and the goodness of fit determines how accurate library quantifications are. Here, we noted that the fit is not optimal: correlation coefficient R^2, which should be over 0.99, is ~ 0.96, and amplification efficiency, which is calculated based on the slope of the standard curve, is higher than expected: 166% and 159% for the second and first library quantification respectively (Table 7). For library quantification, expected efficiency is between 90%–110%, corresponding to an amplification factor of ~ 2. The excess amplification efficiency is an artifact caused by the flat slope of the standard curve, which most likely arises from inaccuracies in standard series dilutions (Figure 14). The same dilutions were used in all qPCRs; thus, the effect is likely to be similar in all qPCR quantification. It should be noted that the overestimate of efficiency also leads to overestimates of the molecule copy numbers in the DNA libraries, suggesting that the library copy numbers presented above may be higher than the true copy numbers.

Table 7: Regression parameter values derived from qPCR standard curves of the first and second library quantification, and their corresponding expected values.

	1ST	2ND	EXPECTED
SLOPE	-2.35	-2.42	-3.58–-3.10
AMPLIFICATION FACTOR	2.66	2.59	~ 2
EFFICIENCY	166%	159%	90–110%
ERROR	0.93	0.93	~ 0
R^2	0.96	0.96	>0.99
Y-INTERCEPT	28.39	29.29	

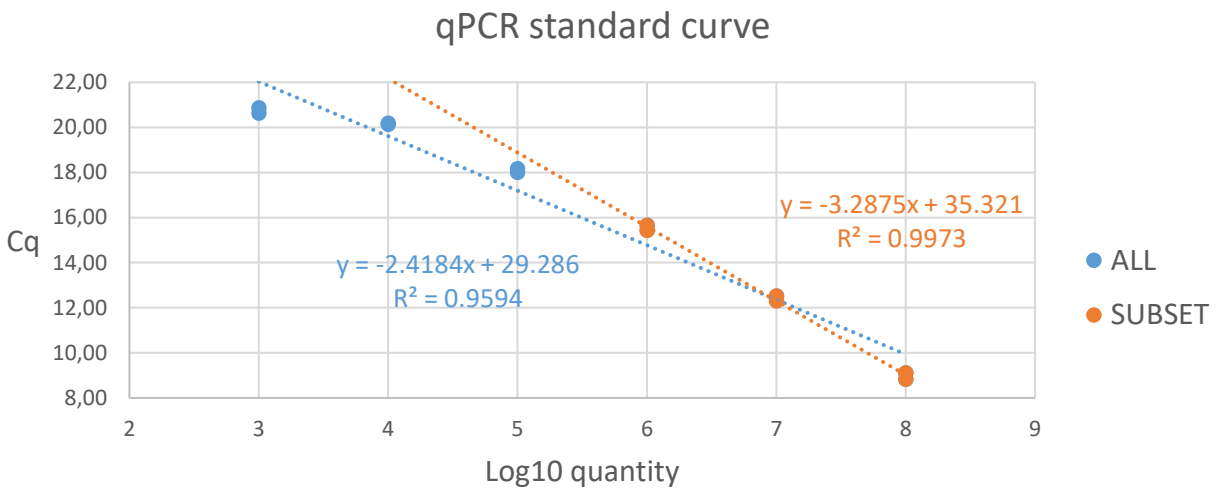

Figure 14: Standard curve from the 2nd library quantification (blue), showing a poor fit of linear regression line. A much better fit is achieved when using only the three standards with highest template concentration (orange), demonstrating that errors in quantification likely arose from inaccurately diluted standards.

After Herculase amplifications, libraries were quantified with a TapeStation system, which is not affected by the above error. The first Herculase amplification yielded concentrations lower than anticipated, yet sufficient for shotgun sequencing (Figure 15A). The two additional rounds of Herculase amplification, which were carried out to increase DNA concentration for mitochondrial enrichment, worked poorly: the average DNA concentration after two rounds of amplification was lower than before amplification (Figure 15B). Eight libraries had lower concentrations, apparently due to DNA loss during purification. Three libraries showed over 2-fold increase and three a slight increase in concentration, and only one of them reach the desired concentration range. Two samples failed quantification due to unknown reasons.

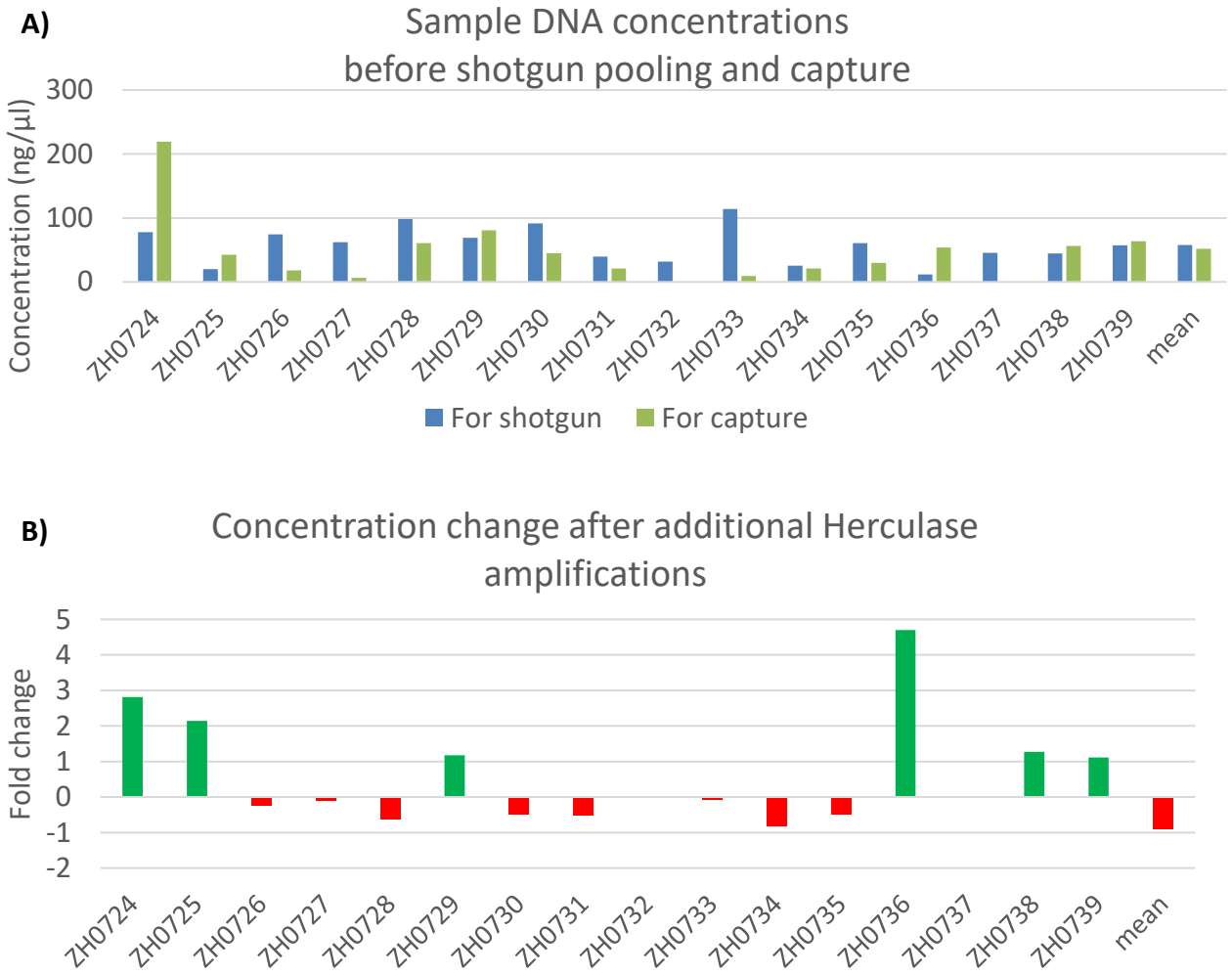

Figure 15: A) Sample DNA concentrations after first Herculase amplification and after the third reamplification. Samples were pooled for shotgun sequencing after the first Herculase amplification and for capture after the third amplification. Final library concentration before capture is missing for samples ZH0732 and ZH0737. B) Fold change of library concentration from first Herculase amplification to third Herculase amplification.

Despite suboptimal DNA concentrations in the second and third enrichment pool, we managed to produce decent capture libraries from all three pools. The pools with lower DNA concentration worked at least equally well when compared to the pool with sufficient concentration; the third, poorest pool had the highest concentration after the final amplification (Table 8). Average fragment length was a little longer than expected for ancient DNA, a possible indication of modern contamination or unspecific enrichment of environmental, non-target sequences.

Table 8: Final DNA concentrations in capture pools after Herculase amplification.

	CONCENTRATION AFTER CAPTURE (COPIES/µL)	CONCENTRATION AFTER FINAL AMPLIFICATION (ng/µL)	AVERAGE SIZE (BP)
POOL 1	2.68E+07	28.7	228
POOL 2	5.48E+07	28.5	238
POOL 3	8.31E+07	73.3	245

4.5 Data quality and contamination

FastQC analysis on raw reads raised no unusual warnings, and the quality of the sequencing reads was considered good. The first quality check indicated a high adapter content, which was expected due to short fragment size. The second FastQC analysis was done on clipped and merged reads, and it indicated that most adapters had been successfully removed (Figure A5.1). The mean length of the merged reads across samples ranged from 61 to 97 bp, demonstrating an effective recovery of relatively short fragments (Table 9). However, the DNA extraction method used here is expected to produce fragment length distribution with a mode below 50 bp (Rohland et al., 2018), but this was achieved only for the Mesolithic lakebed samples and the Iron Age burial sample (Figure A5.2). Stone Age woodland samples had fragment length distributions skewed toward longer fragments.

The average merging rate ranged from 57% to 87%, and from 73% to 91% for shotgun sequenced and enriched samples respectively (Table 9). Lower merge rates likely reflect the presence of longer fragments in the libraries, which is unsurprising when the DNA has been extracted from soil, where environmental sequences are likely to be abundant. The number of reads mapping to 242 mammalian mitochondrial reference genomes after duplicate removal in the shotgun sequenced samples was extremely low, between 0 and 27, further suggesting that the vast majority of the sequenced reads are environmental. In the enriched samples, the corresponding counts were between 6 and 453 reads, showing reasonable enrichment of the target sequences given the very low proportion of mapping reads in the shotgun sequenced libraries. However, the absolute number of mapping reads is still extremely low, even for ancient DNA.

Cluster factor, which indicates the average number of duplicates per sequence and should be close to one, was low in all shotgun sequenced libraries. Enriched samples had overall higher duplication rates than their shotgun sequenced counterparts: half of the enriched libraries had cluster factor between 1 and 2, which is reasonable, while 5 libraries had cluster factor between 2 and 5, indicating that the libraries had been "sequenced to exhaustion" (i.e. no further complexity can be achieved by deeper sequencing). Three enriched samples showed exceptionally high duplication rates: 55 and 68-fold for two Kammarlahti samples and 88 for one Karpankangas sample. High duplication rates possibly arose from the repeated Herculase amplifications, which may have significantly decreased library complexity in some of the samples.

Since negative controls from mitochondrial enrichment were not yet sequenced by the time of writing, I used shotgun sequenced negative controls to estimate laboratory contamination. All

controls had a low number of mapping reads: 0 and 18 in extraction controls and 8 reads in the library control. Because human DNA is usually the most common laboratory contaminant, and thus some sporadic human hits are expected, the small number of mapping reads itself is not worrying (Key et al., 2017). However, library complexity is expected to be lower in negative controls than samples, although detecting the difference often requires slightly deeper sequencing than done here. Nevertheless, the cluster factor in all controls is overall slightly higher than in shotgun sequenced samples, indicating lower library complexity. The first extraction control and the library control have clearly less sequences than the sample libraries, indicating a low contamination rate, while the second extraction control has a read count well within the range of samples.

Table 9: Summary of sequence analysis results after adapter clipping, merging, mapping to 242 mammalian reference set and duplicate removal.

Site	Sample Name	Shotgun /Enrichment	# reads after C&M prior mapping	% merged reads	mean fragment length (bp)	# mapped reads prior RMDup	# mapped reads after RMDup	Cluster Factor
Spångkärret	ZH0724	Shotgun	2692676	78	83	1	1	1.00
		Enrichment	3836008	80	84	19	12	1.58
	ZH0725	Shotgun	14314632	66	99	15	9	1.67
		Enrichment	6370280	73	97	531	377	1.41
Niskasuo	ZH0726	Shotgun	6458799	78	91	4	4	1.00
		Enrichment	4334956	86	89	13	7	1.86
	ZH0727	Shotgun	5585387	63	93	2	2	1.00
		Enrichment	3115475	75	89	45	13	3.46
Karpankangas	ZH0728	Shotgun	5801156	79	84	8	7	1.14
		Enrichment	10510253	82	84	3261	37	88.14
	ZH0729	Shotgun	3790873	79	81	2	2	1.00
		Enrichment	7253691	82	81	89	19	4.68
Konstunkangas	ZH0730	Shotgun	1661405	61	92	1	1	1.00
		Enrichment	10539165	75	90	74	33	2.24
	ZH0731	Shotgun	12595479	75	79	11	10	1.10
		Enrichment	3407129	78	76	27	21	1.29
	ZH0732	Shotgun	17537161	79	81	9	9	1.00
		Enrichment	12850476	84	81	99	55	1.80
Taipaleenranta	ZH0733	Shotgun	2406253	57	103	2	2	1.00
		Enrichment	2751255	74	97	6	6	1.00
	ZH0734	Shotgun	26499586	85	79	10	9	1.11

		Enrichment	4611240	91	76	75	53	1.42
	ZH0735	Shotgun	8028461	84	78	5	4	1.25
		Enrichment	13267302	87	77	256	134	1.91
Kammarlahti	ZH0736	Shotgun	3435064	85	65	4	4	1.00
		Enrichment	12172795	88	67	25104	453	55.42
	ZH0737	Shotgun	8455042	82	59	1	1	1.00
		Enrichment	12684410	85	61	59	17	3.47
	ZH0738	Shotgun	19600110	87	63	28	27	1.04
		Enrichment	13332517	91	63	7765	114	68.11
Toppolanmäki	ZH0739	Shotgun	3832702	80	69	0	0	NA
		Enrichment	8734583	85	69	166	61	2.72
Extraction control	ZH0739EB1	Shotgun	47808	76	76	0	0	NA
	ZH0739EB2	Shotgun	6283579	76	75	25	18	1.39
Library control	ZH0739LB	Shotgun	194844	7	57	11	8	1.38

5 DISCUSSION

Human DNA preservation in the Stone Age sediments

I analysed DNA from archaeological sediment samples collected from six Finnish Stone Age settlement sites and one Iron Age burial. I used targeted enrichment of human mtDNA and high-throughput sequencing to assess the prospects of ancient DNA preservation in Finnish soil. We were able to recover only minute amounts of human DNA from the sediment samples. Human mtDNA was detected in the sediments of three Neolithic settlement sites and a submerged Mesolithic settlement site. Samples from two sites had enough sequences to estimate ancient DNA authenticity: the samples from submerged Kammarlahti had relatively long fragment length and showed no signs of DNA degradation, suggesting that the human mtDNA sequences likely originate from modern contamination.

A sample from the Stone Age Spångkärret site carried shorter mtDNA fragments, which presented a terminal C-to-T substitution pattern characteristic for degraded DNA, albeit at low frequency. Based on ceramics, the time of settlement at the Spångkärret site, and therefore the expected time of DNA deposition, was 5,900–5,200 years ago. The speed of DNA degradation can vary substantially between environments, sites, and even samples, and thus damage frequencies cannot be used to infer sample age. However, a model based on ancient DNA from mammalian bones suggests that the expected deamination frequency for a sample of this age, when preserved

in temperate environment, should be above 20% (Kistler et al., 2017). If the DNA truly originates from the Stone Age, I would expect to see significantly higher damage frequencies than seen here. Thus, the origin of the DNA fragments remains inconclusive: it is possible that the DNA ended up in the soil in later historical times after the Stone Age, but since DNA degradation in soil can be extremely fast, the level of damage seen here might have accumulated in a very short time, within months, weeks or even days. Furthermore, even though the DNA was recovered from the sediment layer that likely dates to the time of the settlement, the horizon structure of the soil is relatively shallow, and it is not possible to exclude DNA leaching from the surface or reposition of the strata.

Another possibility, however, is that the sample from Spångkärret carries both authentic ancient DNA and significant proportion of modern contamination. The mixture of authentic damaged fragments and undamaged modern fragments would lower overall deamination frequencies and complicate authentication (Meyer et al., 2016). Because the number of recovered sequences is very low, it would not require many modern molecules to dilute the damage signal. Yet, with only a handful of sequences, it is extremely difficult to distinguish modern DNA from putatively ancient fragments.

DNA preservation in the Iron Age burial soil

In addition to Stone Age sediments, I also analysed a soil sample from an Iron Age burial to compare DNA preservation between bone and soil under the same environmental conditions. The bones excavated from the grave had previously yielded ancient DNA, and the soil had been collected from the immediate proximity of the remains. However, no detectable amount of human mtDNA was observed in the burial soil sample. This could indicate that DNA is not preserved in the soil even when the environmental conditions favour DNA preservation in the physical remains of the same age. DNA is released into soil during body decomposition (Emmons et al., 2017), but it may be more susceptible to environmental microbes and other decomposing factors than the DNA bound to bone matrix, and thus become more rapidly degraded. Leaching may also play a role here, as water and the fluids released from the decomposing body could carry DNA molecules downwards in the soil. Analysing soil below the burial layer could help to resolve the issue if this is the case.

Challenges from environmental diversity

Ancient samples often contain only a minor proportion of endogenous DNA, while majority of the DNA molecules come from various micro-organism or modern contaminants present in the samples. This metagenomic nature is even more prominent in ancient DNA extracted from soils and sediments and must be considered when processing the data. Even when endogenous ancient DNA is present, it would likely make up only a tiny proportion of all extracted sequences and obtaining enough data to make inferences beyond mere detection of a species would likely require deeper sequencing. Thus, given the substrate and the sequencing depth used in this study, the low number of mapping sequences on human mitochondria is not surprising.

With highly metagenomic data and only a few hundred mapping reads, it is crucial to distinguish true positives from spurious alignments. Taxonomic classification is one way to handle environmental sequencing data. Based on the simulated data, Kraken works well for the purpose: it is fast, flexible and does not have severe downfalls in accuracy. However, the simulated data I used to test the program performance likely poorly represents the real diversity of the environmental DNA composition of Finnish woodland soils: only a fraction of all existing microbial genomes has been sequenced and is available in databases. Moreover, the simulated dataset used here was originally designed to resemble sequence data from cave sediments, which are likely to have a very different microbial composition than woodland soils. Nevertheless, using Kraken to analyse the woodland sediment data provided further insights into program benchmarking: the repetitive sequences that were assigned to Mustelidae suggests that masking repeats and low-complexity regions from the reference genomes before building the database would likely decrease the false discovery rate. Further testings with Kraken could be carried out to refine the estimates of the accuracy. Here, the precision and sensitivity were only calculated for Hominidae family without including all species in the simulated dataset. Additionally, accuracy was only estimated on the family level, and could be different on other taxonomic levels.

Sediments as a source of ancient non-human DNA

The workflow presented here was designed to detect other mammalian species in addition to humans. Although the enrichment predominantly targeted human mtDNA, the probes might have nevertheless picked up mammalian mtDNA fragments as well, since they share plenty of homologous regions with the human mtDNA genome. Traces of animal DNA in the waste pits of

the settlement sites could have been informative on the diet or other animal product usage of the Stone Age people. Yet, no significant hits to other mammalian species were found. Given that the number of observed human sequences was also very small, it is difficult to assess whether the lack of other mammalian sequences is due to their absence in the samples, or human-specific capture probes' inability to enrich them. In this study, mammalian DNA was only a secondary target, and if one wanted to target it more seriously, a better probe design would be needed.

Despite poor human and mammalian DNA preservation, archaeological sediments can have prospects as a source of ancient parasite DNA, since the eggs of some intestinal helminths are naturally persistent in the soil (Morrow et al., 2016). Many parasites require a specific host species and their presence in the latrines or waste pits of prehistoric settlement sites can be used as a marker to infer which animals the people consumed or lived with (e.g. Søe et al., 2018; Tams et al., 2018). Additionally, parasites themselves are informative on the hygiene and health in prehistoric communities. Although previous studies have identified parasite eggs from the sediments under the microscope before sequencing, the shotgun sequenced data produced in this study could also be readily screened for candidate parasites species.

Alternative sampling methods

Obviously, the chosen sampling approach may have affected the results. We collected sediment samples from the five Neolithic Stone Age sites with very little prior knowledge of what lies below the earth surface. We targeted the banks of the house pits in the hope of hitting spots where human excrements and other waste materials might have accumulated, but in practice, we cannot know if this was achieved. Since DNA preservation can be extremely localised and vary substantially even in microscale, it is possible that our sampling simply missed the positions where ancient DNA is present. Additionally, due to limited time and resources, we only analysed a subset of collected sediment samples, further increasing the chance of missing putative ancient DNA. While the unexcavated nature of the Stone Age settlement sites was considered as an advantage, collecting samples during comprehensive excavations might offer valuable information about the underlying structure of the site and help to locate waste pits and other potential areas for sediment DNA sampling. Moreover, instead of extracting DNA from the bulk sediment directly, a more targeted approach could be chosen. For example, sediments could be screened under a microscope to

pinpoint tiny bone fragments, coprolites or parasite eggs. This could also help to enrich endogenous DNA over the overwhelming environmental DNA from the extraction stage onwards.

DNA extraction and inhibitor carry-over

Technical problems in DNA extraction, library preparation and mitochondrial enrichment may have also affected the recovered DNA quantity. The dark extract colour indicated the presence of humic acids or other soil-derived substances in the extracts, and inefficient Herculase amplification suggested that not all inhibitors were removed during purifications. Additionally, inefficient amplification was likely affected by inaccurate qPCR quantification: the number of Herculase amplification cycles we used to reach the required concentration may have been insufficient, because the cycle number was determined based on the qPCR results. This would at least partially explain why we failed to reach the desired library concentrations later on. However, even if that is taken into consideration, DNA quantities were low throughout all laboratory steps. We did not formally monitor inhibition in our samples, but ancient sediments are known to have more inhibitory substances than bones and teeth (Rohland et al., 2018). Therefore, it would be recommended to use better library controls, such as short oligonucleotides that are added to libraries in known quantities, to monitor library conversion efficiency (Glocke and Meyer, 2017).

Testing alternative extraction protocols would be one possible prospect to tackle the presence of inhibitory molecules: for example, a commercial kit designed for faecal materials proved successful in a study of ancient birch bark pitch mastics, which also have high inhibitor content (Kashuba et al., 2019). We used an extraction method that is known to work well for bones, teeth and sediments alike (Rohland et al., 2018), but protocols optimised specifically for sediments also exist, and could be worth trying (e.g. Epp et al. 2019).

Unremoved duplicates

Visualisation of the aligned mtDNA sequences from Kammarlahti revealed that not all duplicates were successfully removed. This was later found to be an artifact produced by the base quality clipping option embedded into AdapterRemoval's default settings, which removes low-quality bases from the 5′ ends of merged reads. This can result in duplicate reads having different coordinates, masking them from duplicate removal. Unremoved duplicates can affect coverage, skew damage frequency estimates, and lead to incorrect variant calls, although the latter was not

done here. The Kammarlahti samples had extremely high duplication rates (> 50), which means that coverage was likely overestimated due to unremoved duplicates. Considering the Spångkärret sample, where cluster factor was reasonably low, unremoved duplicates are likely to have a smaller effect. However, for rigorous results, the analysis should be repeated without base quality clipping.

Contamination

Because negative controls for the mitochondrial enrichment were captured separately and sequenced later, they did not make into this work. Instead, shotgun sequenced controls were used as a proxy for laboratory contamination. Library copy numbers quantified with qPCR indicated low levels of contamination. After sequencing, one of the two extraction controls had a similar number of total and mapping reads as many of the samples, which suggest that laboratory contamination may be present in the samples. However, the second extraction control and the library control indicate low or very low levels of contamination. Additionally, although the number of mapping reads in the negative controls and samples are within the same range, the absolute number of mapping reads is still acceptably low in the controls – the number of human sequences just happens to be unfortunately low in the samples as well.

Modern human contamination in the samples themselves was here assessed only based on C-to-T substitutions and fragment length. More sophisticated means to estimate contamination could not be applied because they require more data. Furthermore, many rely on the assumption that there is only one authentic mtDNA haplotype present in the sample, while others represent modern contamination. This may not be conceptually appropriate for ancient mtDNA extracted from sediments, since the fragments could originate from more than one individual.

6 CONCLUSIONS

Archaeological sediments are a tempting source material for ancient DNA research, since they are abundant, easily accessible, and could provide answers to a wide variety of research questions. Yet, our attempt to recover ancient human mtDNA from Stone Age sediments was largely unsuccessful. A small amount of human mtDNA carrying deamination patterns characteristic for ancient DNA was recovered from one Neolithic site, Spångkärret, but the time of deposition is difficult to estimate. The recovered DNA does not appear damaged enough to come from the Stone Age, but it is impossible to say if it has been deposited one year, 100 years or thousand years ago. Furthermore,

contamination could affect the authenticity estimation, but the small amount of sequences makes contamination difficult to detect.

Interestingly, no human DNA was recovered from the Iron Age burial soil, despite relatively good DNA preservation in the adjacent bones. This observation may suggest that either body decomposition does not release DNA into soil in significant quantities, or that released DNA is quickly degraded.

Thus, based on the samples analysed in this study, the preservation of human DNA in Finnish archaeological sediments may be poor or locally restricted. Even if human DNA was present in the sediments, the overwhelming abundance of environmental sequences complicates its detection, and it would be difficult to obtain enough data to make inferences beyond species detection. Furthermore, DNA extraction from sediments may be hindered by co-extraction of inhibitory substances, such as humic acids. The metagenomic nature of the data exerts some further requirements for computational analyses, in addition to the small amount, short fragment length and damage-derived substitutions typical for ancient DNA.

A simple extraction of human DNA from archaeological sediments attempted by this study does not seem to solve the unfortunate material constraint set by poor bone preservation in Finland. Nevertheless, sediments can still have potential in ancient DNA research and the studies of human history: for example, a more refined sampling approach, targeting microscopic bone fragments or parasite eggs could be a strategy worth trialling. Moreover, given that very few studies concerning archaeological sediments as a source of ancient human DNA has been published, the protocols may not yet be properly established and advancements in the field may provide new possibilities in the future.

Acknowledgements

I would like to thank my supervisors Päivi Onkamo and Verena Schünemann for their guidance and support throughout this project. I also want to thank Petro Pesonen, who contributed to sample collection and site selection, Satu Koivisto and Ulla Moilanen who provided additional samples, and Enrique Rayo and the rest of the IEM Paleogenetics group who guided me through the lab work. In addition, I thank Viviane Slon, Matthias Mayer and Benjamin Vernot for sharing the simulation data and supervising my work with Kraken. I also want to thank my colleagues in the SUGRIGE-project for their collaboration and helpful discussions. Lastly, I thank Jane and Aatos Erkko Foundation, the Kone Foundation and The Kuopio Naturalists' Society for their financial support.

References

Ahola, M., Salo, K., and Mannermaa, K. (2016). Almost gone: Human skeletal material from Finnish Stone Age earth graves. Fennoscandia Archaeol. XXXIII 95–122.

Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature 522, 167–172.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. USA 104, 14616–14621.

Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., et al. (2013). Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. Am. J. Hum. Genet. 93, 852–864.

Crecchio, C., and Stotzky, G. (1998). Binding of DNA on humic acids: Effect on transformation of Bacillus subtilis and resistance to DNase. Soil Biol. Biochem. 30, 1061–1067.

Dabney, J., Meyer, M., and Pääbo, S. (2013a). Ancient DNA damage. Cold Spring Harb. Perspect. Biol. 5, a012567.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., Garcia, N., Paabo, S., Arsuaga, J.-L., et al. (2013b). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. 110, 15758–15763.

Emmons, A.L., DeBruyn, J.M., Mundorff, A.Z., Cobaugh, K.L., and Cabana, G.S. (2017). The persistence of human DNA in soil following surface decomposition. Sci. Justice 57, 341–348.

Epp L.S., Zimmermann, H.H., and Stoof-Leichsenring, K.R. (2019). Sampling and extraction of ancient DNA from sediments in Shapiro, B., Barlow, A., Heintzman, P.D., Hofreiter, M., Paijmans, J.L.A., and Soares, A.E.R. (ed.) Acient DNA: Methods and Protocols (Springer Science+Business Media), pp. 31-44.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048.

de Filippo, C., Meyer, M., and Prüfer, K. (2018). Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences. BMC Biol. 16, 121.

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H. a, Kelso, J., and Pääbo, S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. Proc. Natl. Acad. Sci. USA 110, 2223–2237.

Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., et al. (2015). An early modern human from Romania with a recent Neanderthal ancestor. Nature 524, 216–219.

Fulton, T.L. and Shapiro, B (2019). Setting up an ancient DNA laboratory in Shapiro, B., Barlow, A., Heintzman, P.D., Hofreiter, M., Paijmans, J.L.A. and Soares, A.E.R. (ed.) Acient DNA: Methods and Protocols (Springer Science+Business Media), pp. 1-14.

Furtwängler, A., Reiter, E., Neumann, G.U., Siebke, I., Steuri, N., Hafner, A., Lösch, S., Anthes, N., Schuenemann, V.J., and Krause, J. (2018). Ratio of mitochondrial to nuclear DNA affects contamination estimates in ancient DNA analysis. Sci. Rep. 8, 14075.

Gamba, C., FernÚndez, E., Tirado, M., Deguilloux, M.F., Pemonge, M.H., Utrilla, P., Edo, M., Molist, M., Rasteiro, R., Chikhi, L., et al. (2012). Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. Mol. Ecol. 21, 45–56.

Gansauge, M.-T., and Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat. Protoc. 8, 737–748.

Glocke, I., and Meyer, M. (2017). Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. Genome Res. 27, 1230–1237.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal genome. Science 328, 710–722.

Haak, W., Balanovsky, O., Sanchez, J.J., Koshel, S., Zaporozhchenko, V., Adler, C.J., Der Sarkissian, C.S.I., Brandt, G., Schwarz, C., Nicklisch, N., et al. (2010). Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PLoS Biol. 8, e1000536.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522, 207–211.

Haggren, G., Halinen, P., Lavento, M., Raninen, S., and Wessman, A. (2015). Muinaisuutemme jäljet: Suomen esi- ja varhaishistoria kivikaudelta keskiajalle (Gaudeamus). pp. 55–121.

Haile, J., Holdaway, R., Oliver, K., Bunce, M., Gilbert, M.T.P., Nielsen, R., Munch, K., Ho, S.Y.W., Shapiro, B., and Willerslev, E. (2007). Ancient DNA chronology within sediment deposits: Are paleobiological reconstructions possible and is DNA leaching a factor? Mol. Biol. Evol. 24, 982–989.

Haile, J., Froese, D.G., MacPhee, R.D.E., Roberts, R.G., Arnold, L.J., Reyes, A. V., Rasmussen, M., Nielsen, R., Brook, B.W., Robinson, S., et al. (2009). Ancient DNA reveals late survival of mammoth and horse in interior Alaska. Proc. Natl. Acad. Sci. 106, 22352–22357.

Hofreiter, M., Serre, D., Poinar, H.N., Kuch, M., and Pääbo, S. (2001). Ancient DNA. Nat. Rev. Genet. 2, 353–359.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. 17, 377–386.

Kashuba, N., Kırdök, E., Damlien, H., Manninen, M.A., Nordqvist, B., Persson, P., and Götherström, A. (2019). Ancient DNA from mastics solidifies connection between material culture and genetics of mesolithic hunter–gatherers in Scandinavia. Commun. Biol. 2, 185.

Key, F.M., Posth, C., Krause, J., Herbig, A., and Bos, K.I. (2017). Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. Trends Genet. 33, 508–520.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40, e3.

Kistler, L., Ware, R., Smith, O., Collins, M., and Allaby, R.G. (2017). A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Res. 45, 6310–6320.

Krause, J., Fu, Q., Good, J.M., Viola, B., Shunkov, M. V, Derevianko, A.P., and Pääbo, S. (2010). The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 464, 894–897.

Lamnidis, T.C., Majander, K., Jeong, C., Salmela, E., Wessman, A., Moiseyev, V., Khartanovich, V., Balanovsky, O., Ongyerth, M., Weihmann, A., et al. (2018). Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. Nat. Commun. 9, 5018.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513, 409–413.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595.
Lindhal, T. (1993). Instability and decay of the primary structure of DNA. Nature 362, 709–715.
Massilani, D., Aldeias, V., Miller, C., Morley, M., Goldberg, P., Slon, V., and Meyer, M. (2018). Ancient DNA in sediment – a micromorphology approach. In 8th International Symposium on Biomolecular Archaeology ISBA 2018, (Jena, Germany).
Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 2010, 10.1101/pdb.prot5448.
Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222–226.
Meyer, M., Fu, Q., Aximu-Petri, A., Glocke, I., Nickel, B., Arsuaga, J.-L., Martínez, I., Gracia, A., de Castro, J.M.B., Carbonell, E., et al. (2014). A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 505, 403–406.
Meyer, M., Arsuaga, J.L., De Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martínez, I., Gracia, A., De Castro, J.M.B., Carbonell, E., et al. (2016). Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature 531, 504–507.
Moilanen, U. (2017). Valkeakoski, Toppolanmäki: Haudan nro 3/1937 tutkimuskaivaus. Available at request from the Finnish Heritage Agency.
Morrow, J.J., Newby, J., Piombino-Mascali, D., and Reinhard, K.J. (2016). Taphonomic considerations for the analysis of parasites in archaeological materials. Int. J. Paleopathol. 13, 56–64.
Ogram, A. V, Mathot, M.L., Harsh, J.B., Boyle, J., and Pettigrew, C.A. (1994). Effects of DNA polymer length on its adsorption to soils. Appl. Environ. Microbiol. 60, 393–396.
Overballe-Petersen, S., Orlando, L., and Willerslev, E. (2012). Next-generation sequencing offers new insights into DNA degradation. Trends Biotechnol. 30, 364–368.
Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., and Hofreiter, M. (2004). Genetic analyses from ancient DNA. Annu. Rev. Genet. 38, 645–679.
Pedersen, M.W., Overballe-Petersen, S.S., Ermini, L., Sarkissian, C. Der, Haile, J., Hellstrom, M., Spens, J., Thomsen, P.F., Bohmann, K., Cappellini, E., et al. (2015). Ancient and modern environmental DNA. Philos. Trans. R. Soc. B-Biological Sci. 370, 20130383.
Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016). EAGER: efficient ancient genome reconstruction. Genome Biol. 17, 60.
Pesonen, P. (2018). KOTKA, LOVIISA, TAIPALSAARI, VIROLAHTI, Näytteenotto kivikautisista asuinpaikoista 2018. Tutkimusraportti. Available online at: https://asiat.museovirasto.fi/case/MV/181/05.04.01.00/2018.
Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M.R., Pugach, I., Ko, A.M.-S., Ko, Y.-C., Jinam, T.A., Phipps, M.E., et al. (2011). Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am. J. Hum. Genet. 89, 516–528.
Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. 29, 24–26.
Rohland, N., Glocke, I., Aximu-Petri, A., and Meyer, M. (2018b). Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nat. Protoc. 13, 2447–2461.

Sanchez-Quinto, F., and Lalueza-Fox, C. (2015). Almost 20 years of Neanderthal palaeogenetics: adaptation, admixture, diversity, demography and extinction. Philos. Trans. R. Soc. B Biol. Sci. 370, 20130374.

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88.

Skoglund, P., Malmstrom, H., Raghavan, M., Stora, J., Hall, P., Willerslev, E., Gilbert, M.T.P., Gotherstrom, A., and Jakobsson, M. (2012). Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. Science 336, 466–469.

Skoglund, P., Northoff, B.H., Shunkov, M. V, Derevianko, A.P., Pääbo, S., Krause, J., and Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc. Natl. Acad. Sci. 111, 2229–2234.

Slon, V., Glocke, I., Barkai, R., Gopher, A., Hershkovitz, I., and Meyer, M. (2016). Mammalian mitochondrial capture, a tool for rapid screening of DNA preservation in faunal and undiagnostic remains, and its application to Middle Pleistocene specimens from Qesem Cave (Israel). Quat. Int. 398, 210–218.

Slon, V., Hopfe, C., Weiß, C.L., Mafessoni, F., de la Rasilla, M., Lalueza-Fox, C., Rosas, A., Soressi, M., Knul, M. V., Miller, R., et al. (2017). Neandertal and Denisovan DNA from Pleistocene sediments. Science 356, 605–608.

Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., Hajdinjak, M., Peyrégne, S., Nagel, S., Brown, S., et al. (2018). The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature 561, 113–116.

Smith, O., Momber, G., Bates, R., Garwood, P., Fitch, S., Pallen, M., Gaffney, V., and Allaby, R.G. (2015). Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago. Science 347, 998–1001.

Smith, T.F., Waterman, M.S., and Burks, C. (1985). The statistical distribution of nucleic acid similarities. Nucleic Acids Res. 13, 645–656.

Søe, M.J., Nejsum, P., Seersholm, F.V., Fredensborg, B.L., Habraken, R., Haase, K., Hald, M.M., Simonsen, R., Højlund, F., Blanke, L., et al. (2018). Ancient DNA from latrines in Northern Europe and the Middle East (500 BC–1700 AD) reveals past parasites and diet. PLoS One 13, e0195481.

Tams, K.W., Søe, M.J., Merkyte, I., Seersholm, F.V., Henriksen, P.S., Klingenberg, S., Willerslev, E., Kjær, K.H., Hansen, A.J., and Outzen Kapel, C.M. (2018). Parasitic infections and resource economy of Danish Iron Age settlement through ancient DNA sequencing. PLoS One 13, e0197399.

van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L., and Guschanski, K. (2019). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol. Ecol. Resour. doi: 10.1111/1755-0998.13009.

Willerslev, E., Hansen, A.J., Binladen, J., Brand, T.B., Gilbert, M.T.P., Shapiro, B., Bunce, M., Wiuf, C., Gilichinsky, D.A., and Cooper, A. (2003). Diverse Plant and Animal Genetic Records from Holocene and Pleistocene Sediments. Science 300, 791–795.

Willerslev, E., Cappellini, E., Boomsma, W., Nielsen, R., Hebsgaard, M.B., Brand, T.B., Hofreiter, M., Bunce, M., Poinar, H.N., Dahl-Jensen, D., et al. (2007). Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 317, 111–114.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46.

Appendices

Appendix 1: DNA library preparation reaction master mixes.

Table A1.1: Blunt-end repair master mix

ADDED VOLUME	REAGENT
472µL	DI water
100 µL	10x NEBuffer2
100 µL	10 mM ATP
40 µL	20 mg/ml BSA
40 µL	2.5 mM dNTPs
40 µL	T4 PNK
8 µL	T4 DNA polymerase

Table A1.2: Fill-in master mix

ADDED VOLUME	REAGENT
240 µL	DI water
80 µL	Isothermopol buffer (10X)
40 µL	2.5 mM dNTPs
40 µL	8 U/µL Bst polymerase

Table A1.3: Indexing master mix

ADDED VOLUME	REAGENT
5920 µL	DI water
800	10X Pfu Turbo Buffer
80 µL	25 mM dNTP mix
60 µL	20 ng/ml BSA
80 µL	Pfu Turbo Polymerase

Table A1.4: Herculase reamplification master mix.

ADDED VOLUME	REAGENT
10868 µL	DI water
3344 µL	5X Herculase II Reaction Buffer
689 µL	10 µM IS5
689 µL	10 µM IS5
167 µL	dNTPs (100 mM, 25mM of each dNTPs)
167 µL	Herculase II Fusion DNA Polymerase

Appendix 2: Library index combinations and qPCR primers.

Table A2.1: List of index combination used for DNA libraries.

SAMPLE	P7 PRIMER	P7 SEQUENCE	P5 PRIMER	P5 SEQUENCE
ZH0724	P7_Tue_8nt_0015	ACCAGACC	P5_Tue_8nt_0008	AACGAAGT
ZH0725	P7_Tue_8nt_0023	ATGAATCT	P5_Tue_8nt_0012	AAGAAGAC
ZH0726	P7_Tue_8nt_0024	ATGCCGGA	P5_Tue_8nt_0023	AGGTTGAC
ZH0727	P7_Tue_8nt_0075	GTTGACCG	P5_Tue_8nt_0024	AGTTGGTT
ZH0728	P7_Tue_8nt_0041	CCTGGTAC	P5_Tue_8nt_0035	CCGCTACG
ZH0729	P7_Tue_8nt_0092	TGGTTGAC	P5_Tue_8nt_0041	CCTTACTG
ZH0730	P7_Tue_8nt_0095	TTCGATGA	P5_Tue_8nt_0051	CTGGTAGG
ZH0731	P7_Tue_8nt_0006	AACCTCAG	P5_Tue_8nt_0054	GAAGAGGT
ZH0732	P7_Tue_8nt_0045	CGTATTGC	P5_Tue_8nt_0058	GCTGCGGA
ZH0733	P7_Tue_8nt_0014	ACCAAGAT	P5_Tue_8nt_0059	GCTTCTTA
ZH0734	P7_Tue_8nt_0039	CCTGACGG	P5_Tue_8nt_0063	GGCTTGCT
ZH0735	P7_Tue_8nt_0071	GTCGCTAG	P5_Tue_8nt_0073	GTTGAATT
ZH0736	P7_Tue_8nt_0036	CCGTCTGC	P5_Tue_8nt_0075	GTTGCTGG
ZH0737	P7_Tue_8nt_0020	AGAGAACG	P5_Tue_8nt_0078	TAGTCTAC
ZH0738	P7_Tue_8nt_0094	TTCAGCAG	P5_Tue_8nt_0082	TCGGTCTC
ZH0739	P7_Tue_8nt_0029	CATCAGTT	P5_Tue_8nt_0087	TGATGAGA
ZH0739EB1	P7_Tue_8nt_0008	AACGAAGT	P5_Tue_8nt_0044	CGAGAGTT
ZH0739EB2	P7_Tue_8nt_0002	AACCAGAA	P5_Tue_8nt_0065	GGTCAGCA
ZH0739LB	P7_Tue_8nt_0032	CCAAGTCA	P5_Tue_8nt_0004	AACCGAAC

Table A2.2: List of primers used in the experiment.

PRIMER	SEQUENCE
IS7	5’-ACACTCTTTCCCTACACGAC
IS8	5’-GTGACTGGAGTTCAGACGTGT
IS5	5’-AATGATACGGCACCACCGA
IS6	5’-CAAGCAGAAGACGGCATACGA

Appendix 3: Accession IDs and names of the 791 mammalian mitochondrial genomes included in the Kraken database.

NC_028718.1	Microcebus murinus
NC_028715.1	Niviventer fulvescens
NC_028592.1	Cercocebus atys
NC_028625.1	Castor fiber
NC_028577.1	Zaedyus pichiy
NC_028576.1	Tolypeutes tricinctus
NC_028575.1	Tolypeutes matacus
NC_028574.1	Tamandua mexicana
NC_028573.1	Priodontes maximus
NC_028572.1	Myrmecophaga tridactyla
NC_028571.1	Euphractus sexcinctus
NC_028570.1	Dasypus yepesi
NC_028569.1	Dasypus septemcinctus
NC_028568.1	Dasypus sabanicola
NC_028567.1	Dasypus pilosus
NC_028566.1	Dasypus kappleri
NC_028565.1	Dasypus hybridus
NC_028564.1	Cyclopes didactylus
NC_028563.1	Chlamyphorus truncatus
NC_028562.1	Chaetophractus villosus
NC_028561.1	Chaetophractus vellerosus
NC_028560.1	Calyptophractus retusus
NC_028559.1	Cabassous unicinctus
NC_028558.1	Cabassous tatouay
NC_028557.1	Cabassous chacoensis
NC_028556.1	Cabassous centralis
NC_028555.1	Bradypus torquatus
NC_028554.1	Bradypus pygmaeus
NC_028536.1	Rhinolophus rex
NC_028501.1	Bradypus variegatus
NC_028442.1	Mandrillus leucophaeus
NC_028427.1	Lycaon pictus
NC_026064.1	Ovis vignei breed Urial
NC_026063.1	Ovis orientalis breed Asian mouflon
NC_028312.1	Prionailurus planiceps
NC_028335.1	Bandicota indica
NC_028323.1	Otocolobus manul
NC_028322.1	Leopardus jacobita
NC_028321.1	Leopardus guigna
NC_028320.1	Leopardus geoffroyi
NC_028319.1	Lynx pardinus
NC_028318.1	Leopardus wiedii
NC_028317.1	Leopardus tigrinus
NC_028316.1	Leptailurus serval
NC_028315.1	Leopardus pardalis
NC_028314.1	Leopardus colocolo
NC_028313.1	Lynx canadensis
NC_028311.1	Puma yagouaroundi
NC_028310.1	Felis silvestris
NC_028309.1	Felis nigripes
NC_028308.1	Felis margarita
NC_028307.1	Felis chaus
NC_028306.1	Caracal caracal
NC_028305.1	Prionailurus viverrinus
NC_028304.1	Prionailurus rubiginosus
NC_028303.1	Pardofelis marmorata
NC_028302.1	Panthera leo
NC_028301.1	Prionailurus bengalensis
NC_028300.1	Catopuma badia
NC_028299.1	Profelis aurata
NC_028210.1	Propithecus verreauxi
NC_028161.1	Capra aegagrus
NC_028150.1	Uropsilus sp. 4 FT-2015
NC_028013.1	Mustela eversmannii
NC_027977.1	Plecotus macrobullaris
NC_027973.1	Myotis petax
NC_027964.1	Choloepus hoffmanni
NC_027963.1	Sorex araneus
NC_027956.1	Canis anthus
NC_027945.1	Microtus ochrogaster
NC_027935.1	Vulpes ferrilata
NC_027932.1	Micromys minutus
NC_027828.1	Mungotictis decemlineata
NC_027825.1	Alouatta seniculus
NC_027742.1	Fukomys damarensis
NC_027740.1	Propithecus tattersalli
NC_026204.2	Crocidura attenuata
NC_027692.1	Stylodipus telum
NC_027684.1	Meriones meridianus
NC_027683.1	Meriones libycus
NC_027658.1	Callithrix kuhlii
NC_026781.1	Manis javanica
NC_026780.1	Manis tricuspis
NC_027604.1	Macaca leonina
NC_027593.1	Mesoplodon ginkgodens
NC_027579.1	Sicista concolor
NC_027578.1	Eozapus setchuanus
NC_027500.1	Euchoreutes naso
NC_027499.1	Dipus sagitta
NC_027449.1	Macaca cyclopis
NC_005943.1	Macaca mulatta
NC_024820.1	Giraffa camelopardalis
NC_006893.1	Crocidura russula
NC_027418.1	Eothenomys melanogaster
NC_027283.1	Spermophilus dauricus
NC_027278.1	Ictidomys tridecemlineatus
NC_027249.1	Crocidura beatus
NC_027248.1	Crocidura mindorus
NC_027247.1	Crocidura grayi
NC_027246.1	Crocidura panayensis
NC_027245.1	Crocidura negrina
NC_027243.1	Crocidura palawanensis
NC_027242.1	Crocidura orientalis
NC_027237.1	Nyctalus noctula
NC_027244.1	Crocidura ninoyi
NC_027233.1	Bison priscus
NC_026456.1	Neophocaena asiaeorientalis
NC_025271.1	Capreolus pygargus
NC_027115.1	Catopuma temminckii
NC_027083.1	Lynx lynx
NC_026992.1	Sus barbatus
NC_026976.1	Macaca nemestrina
NC_026915.1	Myospalax aspalax
NC_026875.1	Tamiops swinhoei
NC_026723.1	Urocyon cinereoargenteus
NC_026706.1	Cynomys ludovicianus
NC_026705.1	Cynomys leucurus
NC_023543.1	Damaliscus lunatus
NC_023542.1	Beatragus hunteri
NC_026542.1	Pteropus vampyrus
NC_026529.1	Vulpes lagopus
NC_026465.1	Cynopterus brachyotis
NC_026460.1	Rhinolophus macrotis
NC_026443.1	Hylopetes phayrei
NC_026442.1	Dremomys rufigenis
NC_020432.2	Equus grevyi
NC_026131.1	Episoriculus caudatus
NC_026124.1	Rhizomys sinensis
NC_026120.1	Macaca nigra
NC_026098.1	Eulemur rubriventer
NC_026095.1	Indri indri
NC_026090.1	Palaeopropithecus ingens
NC_026088.1	Megaladapis edwardsi
NC_026086.1	Propithecus edwardsi
NC_026085.1	Varecia variegata
NC_026084.1	Propithecus diadema
NC_026034.1	Myospalax psilurus
NC_019578.2	Globicephala macrorhynchus
NC_025952.1	Mus spretus
NC_025949.1	Murina leucogaster
NC_025902.1	Lepus hainanus
NC_025777.1	Scapanulus oweni
NC_025769.1	Manis temminckii
NC_025748.1	Lepus tolai
NC_025747.1	Wiedomys cerradensis
NC_025746.1	Akodon montensis
NC_025670.1	Leopoldamys edwardsi
NC_025586.1	Callithrix jacchus
NC_025568.1	Myotis davidii
NC_025563.1	Bos mutus
NC_025559.1	Neomys fodiens
NC_025550.1	Callosciurus erythraeus
NC_025516.1	Mustela erminea
NC_025513.1	Macaca fuscata
NC_025330.1	Cricetulus longicaudatus
NC_025327.1	Sorex tundrensis
NC_025316.1	Lepus sinensis
NC_025308.1	Myotis brandtii
NC_025296.1	Viverricula indica
NC_025287.1	Mus fragilicauda
NC_025283.1	Lasiopodomys mandarinus
NC_025278.1	Sorex cylindricauda
NC_025277.1	Tamias sibiricus
NC_025270.1	Mus cookii
NC_025269.1	Mus cervicolor
NC_025268.1	Mus caroli
NC_012920.1	Homo sapiens
NC_005089.1	Mus musculus
NC_024812.1	Mazama nemorivaga
NC_025222.1	Macaca tonkeana
NC_025221.1	Macaca silenus
NC_025201.1	Macaca arctoides
NC_024942.1	Mustela nigripes
NC_024933.1	Chlorocebus cynosuros
NC_024860.1	Sus celebensis
NC_024819.1	Dama mesopotamica
NC_024818.1	Bos gaurus
NC_011137.1	Homo sapiens neanderthalensis
NC_023536.1	Sus verrucosus
NC_023351.1	Nectogale elegans
NC_021751.1	Mustela altaica
NC_021749.1	Martes martes
NC_021478.1	Rhizomys pruinosus
NC_020006.2	Papio anubis
NC_018598.1	Uropsilus sp. 1 FT-2014
NC_024630.1	Callicebus lugens
NC_024629.1	Chiropotes israelita
NC_024628.1	Callimico goeldii
NC_024604.1	Suncus murinus
NC_024592.1	Cricetulus kamensis
NC_024569.1	Prionodon pardicolor
NC_024568.1	Genetta servalina
NC_024567.1	Nandinia binotata
NC_024563.1	Anourosorex squamipes
NC_024558.1	Vespertilio sinensis
NC_024538.1	Myodes glareolus
NC_024529.1	Trachypithecus pileatus
NC_024259.1	Lepus coreanus
NC_024043.1	Lepus americanus
NC_024042.1	Lepus granatensis
NC_024041.1	Lepus townsendii
NC_024040.1	Lepus timidus
NC_018367.1	Marmota himalayana
NC_024172.1	Chrysocyon brachyurus
NC_024053.1	Tarsius wallacei
NC_024052.1	Tarsius dentatus
NC_024051.1	Tarsius lariang
NC_018781.1	Equus burchellii chapmani
NC_018783.1	Equus ovodovi
NC_018782.1	Equus hemionus kulan
NC_018780.1	Equus zebra hartmannae
NC_024030.1	Equus przewalskii
NC_023922.1	Petaurista alborufus
NC_023971.1	Trachypithecus cristatus
NC_023970.1	Trachypithecus francoisi
NC_023965.1	Allenopithecus nigroviridis
NC_023964.1	Cercocebus torquatus
NC_023963.1	Cercopithecus diana
NC_023962.1	Cercopithecus lhoesti
NC_023961.1	Cercopithecus mitis
NC_023960.1	Niviventer confucianus
NC_023958.1	Vulpes corsac
NC_023950.1	Blarinella quadraticauda
NC_016178.1	Cervus nippon kopschi
NC_015828.1	Myotis formosus
NC_023830.1	Mesoplodon grayi
NC_023541.1	Sus cebifrons
NC_023838.1	Tapirus indicus
NC_023889.1	Orcinus orca
NC_020758.1	Acomys cahirinus
NC_023795.1	Macaca assamensis
NC_023780.1	Ratufa bicolor
NC_023459.1	Vulpes zerda
NC_023457.1	Capricornis milneedwardsii
NC_023347.1	Rattus niobe
NC_023368.1	Pteronotus parnellii
NC_023100.1	Homo heidelbergensis
NC_020657.1	Proechimys longicaudatus
NC_020659.1	Ctenomys leucodon
NC_006914.1	Mus musculus domesticus
NC_023263.1	Meriones unguiculatus
NC_023244.1	Uropsilus soricipes
NC_023211.1	Saimiri oerstedii citrinellus
NC_023210.1	Mustela kathiah
NC_023122.1	Pteropus alecto
NC_023089.1	Petaurista hainana
NC_022842.1	Panthera onca
NC_022805.1	Tursiops australis
NC_022698.1	Myotis ikonnikovi
NC_022694.1	Myotis macrodactylus
NC_022474.1	Eptesicus serotinus
NC_022429.1	Vampyrum spectrum
NC_022428.1	Tonatia saurophila
NC_022427.1	Sturnira tildae
NC_022426.1	Rhinophylla pumilio
NC_022425.1	Pteronotus rubiginosus
NC_022424.1	Lophostoma silvicolum
NC_022423.1	Desmodus rotundus
NC_022422.1	Carollia perspicillata
NC_022421.1	Brachyphylla cavernarum
NC_022420.1	Anoura caudifer
NC_022419.1	Micronycteris megalotis
NC_000845.1	Sus scrofa
NC_020093.1	Moschus chrysogaster
NC_021966.1	Saimiri boliviensis
NC_021957.1	Nomascus leucogenys
NC_021974.1	Mesoplodon densirostris
NC_021967.1	Cacajao calvus
NC_021965.1	Callicebus cupreus
NC_021961.1	Cebus xanthosternos
NC_021960.1	Saguinus oedipus
NC_021959.1	Prolemur simus
NC_021958.1	Nycticebus bengalensis
NC_021956.1	Mandrillus sphinx
NC_021955.1	Loris lydekkerianus
NC_021954.1	Lophocebus aterrimus
NC_021953.1	Lepilemur ruficaudatus
NC_021952.1	Leontopithecus rosalia
NC_021951.1	Lagothrix lagotricha
NC_021950.1	Hapalemur griseus
NC_021949.1	Galago moholi
NC_021948.1	Eulemur rufus
NC_021947.1	Erythrocebus patas
NC_021946.1	Chiropotes albinasus
NC_021945.1	Cheirogaleus medius
NC_021944.1	Cercopithecus albogularis
NC_021943.1	Cercocebus chrysogaster
NC_021942.1	Callithrix pygmaea
NC_021941.1	Callithrix geoffroyi
NC_021940.1	Avahi laniger
NC_021939.1	Aotus azarai

NC_021938.1	Alouatta caraya
NC_021119.1	Murina ussuriensis
NC_021387.1	Coendou insidiosus
NC_021381.1	Naemorhedus goral
NC_021461.1	Neophocaena phocaenoides
NC_021435.1	Ziphius cavirostris
NC_021434.1	Mesoplodon europaeus
NC_021398.1	Crocidura shantungensis
NC_021386.1	Chinchilla lanigera
NC_021129.1	Eospalax cansus
NC_020792.1	Tympanoctomys barrerae
NC_020661.1	Octodon degus
NC_020660.1	Spalacopus cyanus
NC_020658.1	Ctenomys sociabilis
NC_020794.1	Redunca arundinum
NC_020793.1	Oryx beisa
NC_020755.1	Nannospalax judaei
NC_020753.1	Tragulus kanchil
NC_020752.1	Tragelaphus strepsiceros
NC_020751.1	Tragelaphus scriptus
NC_020750.1	Tragelaphus oryx
NC_020749.1	Tragelaphus eurycerus
NC_020748.1	Tragelaphus angasii
NC_020789.1	Rupicapra pyrenaica
NC_020788.1	Tetracerus quadricornis
NC_020754.1	Nannospalax galili
NC_020756.1	Spalax carmeli
NC_020667.1	Simias concolor
NC_020735.1	Philantomba maxwellii
NC_020728.1	Neotragus moschatus
NC_020723.1	Naemorhedus griseus
NC_020719.1	Mazama americana
NC_020714.1	Hyemoschus aquaticus
NC_020683.1	Capra caucasica
NC_020678.1	Antidorcas marsupialis
NC_020759.1	Loxodonta cyclotis
NC_020739.1	Pudu mephistophiles
NC_020731.1	Oreotragus oreotragus
NC_020730.1	Okapia johnstoni
NC_020685.1	Cephalophus adersi
NC_020680.1	Axis axis
NC_020666.1	Procolobus verus
NC_020617.1	Syncerus caffer
NC_020766.1	Ozotoceros bezoarticus
NC_020757.1	Nannospalax golani
NC_020747.1	Sylvicapra grimmia
NC_020746.1	Saiga tatarica
NC_020745.1	Rusa timorensis
NC_020744.1	Rusa alfredi
NC_020743.1	Rucervus duvaucelii
NC_020742.1	Redunca fulvorufula
NC_020741.1	Raphicerus campestris
NC_020740.1	Pudu puda
NC_020738.1	Procapra gutturosa
NC_020737.1	Potamochoerus porcus
NC_020736.1	Philantomba monticola
NC_020734.1	Pelea capreolus
NC_020733.1	Ourebia ourebi
NC_020732.1	Oryx leucoryx
NC_020729.1	Odocoileus hemionus
NC_020727.1	Neotragus batesi
NC_020726.1	Nanger soemmerringii
NC_020725.1	Nanger granti
NC_020724.1	Nanger dama
NC_020722.1	Naemorhedus baileyi
NC_020721.1	Mazama rufina
NC_020720.1	Mazama gouazoupira
NC_020718.1	Madoqua saltiana
NC_020717.1	Madoqua kirkii
NC_020716.1	Litocranius walleri
NC_020715.1	Kobus ellipsiprymnus
NC_020713.1	Hippotragus niger
NC_020712.1	Hippotragus equinus
NC_020711.1	Hippocamelus antisensis
NC_020710.1	Gazella subgutturosa
NC_020709.1	Gazella spekei
NC_020708.1	Gazella leptoceros
NC_020707.1	Gazella gazella
NC_020706.1	Gazella erlangeri
NC_020705.1	Gazella dorcas
NC_020704.1	Gazella cuvieri
NC_020703.1	Gazella bennettii
NC_020702.1	Eudorcas rufifrons
NC_020701.1	Dorcatragus megalotis
NC_020700.1	Dama dama
NC_020699.1	Connochaetes taurinus
NC_020698.1	Connochaetes gnou
NC_020697.1	Hexaprotodon liberiensis
NC_020696.1	Cephalorhynchus heavisidii
NC_020695.1	Cephalophus spadix
NC_020694.1	Cephalophus silvicultor
NC_020693.1	Cephalophus rufilatus
NC_020692.1	Cephalophus ogilbyi
NC_020691.1	Cephalophus nigrifrons
NC_020690.1	Cephalophus natalensis
NC_020689.1	Cephalophus leucogaster
NC_020688.1	Cephalophus jentinki
NC_020687.1	Cephalophus dorsalis
NC_020686.1	Cephalophus callipygus
NC_020684.1	Capreolus capreolus
NC_020682.1	Blastocerus dichotomus
NC_020681.1	Axis porcinus
NC_020679.1	Antilocapra americana
NC_020677.1	Alces alces
NC_020676.1	Alcelaphus buselaphus
NC_020675.1	Aepyceros melampus
NC_020674.1	Addax nasomaculatus
NC_020670.1	Crocuta crocuta
NC_020669.1	Hyaena hyaena
NC_020664.1	Martes pennanti
NC_020656.1	Ovis ammon
NC_020648.1	Mephitis mephitis
NC_020647.1	Nasua nasua
NC_020646.1	Taxidea taxus
NC_020645.1	Arctonyx collaris
NC_020644.1	Melogale moschata
NC_020643.1	Martes foina
NC_020642.1	Martes americana
NC_020641.1	Neovison vison
NC_020640.1	Mustela frenata
NC_020639.1	Mustela nivalis
NC_020638.1	Mustela putorius
NC_020637.1	Mustela sibirica
NC_020633.1	Rupicapra rupicapra
NC_020632.1	Pseudois nayaur
NC_020631.1	Ovibos moschatus
NC_020630.1	Oreamnos americanus
NC_020629.1	Capricornis sumatraensis
NC_020628.1	Hemitragus jemlahicus
NC_020627.1	Damaliscus pygargus
NC_020626.1	Capra sibirica
NC_020625.1	Capra pyrenaica
NC_020624.1	Capra nubiana
NC_020623.1	Capra ibex
NC_020622.1	Capra falconeri
NC_020621.1	Hemitragus jayakari
NC_020620.1	Tragelaphus spekii
NC_020619.1	Tragelaphus imberbis
NC_020618.1	Taurotragus derbianus
NC_020616.1	Pseudoryx nghetinhensis
NC_020615.1	Bubalus depressicornis
NC_020614.1	Boselaphus tragocamelus
NC_020476.1	Equus zebra
NC_020010.2	Papio ursinus
NC_020009.2	Papio papio
NC_020008.2	Papio kindae
NC_020007.2	Papio cynocephalus
NC_020326.1	Rhinolophus ferrumequinum quelpartis
NC_019612.1	Pteromys volans
NC_020433.1	Equus kiang
NC_020017.1	Moschus anhuiensis
NC_016421.1	Oryx dammah
NC_019617.1	Niviventer excelsior
NC_019626.1	Neotetracus sinensis
NC_019802.1	Theropithecus gelada
NC_019801.1	Callicebus donacophilus
NC_019800.1	Ateles belzebuth
NC_019799.1	Aotus lemurinus
NC_019591.1	Orcaella heinsohni
NC_019590.1	Orcaella brevirostris
NC_019589.1	Peponocephala electra
NC_019588.1	Feresa attenuata
NC_019585.1	Apodemus latronum
NC_019584.1	Apodemus draco
NC_019583.1	Trachypithecus johnii
NC_019582.1	Trachypithecus vetulus
NC_019581.1	Trachypithecus shortridgei
NC_019580.1	Trachypithecus germaini
NC_019579.1	Trachypithecus hatinhensis
NC_019577.1	Pseudorca crassidens
NC_019441.1	Globicephala melas
NC_018753.1	Nomascus gabriellae
NC_018603.1	Kobus leche
NC_018595.1	Cervus nippon sichuanicus
NC_018540.1	Hipposideros armiger
NC_018539.1	Rhinolophus luctus
NC_018535.1	Eospalax rothschildi
NC_018098.1	Eospalax baileyi
NC_018096.1	Saimiri boliviensis boliviensis
NC_018063.1	Pygathrix cinerea 2 RL-2012
NC_018062.1	Pygathrix cinerea 1 RL-2012
NC_018061.1	Pygathrix nigripes
NC_018060.1	Rhinopithecus bieti 2 RL-2012
NC_018059.1	Rhinopithecus strykeri
NC_018058.1	Rhinopithecus bieti 1 RL-2012
NC_018057.1	Rhinopithecus brelichi
NC_018053.1	Panthera leo persica
NC_018032.1	Hydropotes inermis argyropus
NC_017599.1	Apodemus chevrieri
NC_016920.1	Muntiacus vuquangensis
NC_016873.1	Lasiurus borealis
NC_016872.1	Plecotus rafinesquii
NC_016871.1	Artibeus lituratus
NC_016707.1	Przewalskium albirostris
NC_016689.1	Pseudois schaeferi
NC_016662.1	Apodemus chejuensis
NC_016470.1	Puma concolor
NC_016428.1	Apodemus agrarius
NC_016427.1	Myodes regulus
NC_016422.1	Oryx gazella
NC_016189.1	Prionailurus bengalensis euptilurus
NC_016191.1	Rhinolophus ferrumequinum korai
NC_016061.1	Equus hemionus
NC_016060.1	Apodemus peninsulae
NC_016055.1	Neodon irene
NC_016008.1	Manis pentadactyla
NC_015889.1	Ovis canadensis
NC_015841.1	Lepus capensis
NC_015529.1	Mammuthus columbi
NC_015486.1	Rhinopithecus bieti
NC_015485.1	Rhinopithecus avunculus
NC_015484.1	Plecotus auritus
NC_015247.1	Odocoileus virginianus
NC_015243.1	Microtus fortis calamorum
NC_015241.1	Microtus fortis fortis
NC_015112.1	Heterocephalus glaber
NC_014855.1	Rattus leucopus
NC_014871.1	Rattus sordidus
NC_014875.1	Procapra przewalskii
NC_014867.1	Rattus fuscipes
NC_014864.1	Rattus villosissimus
NC_014861.1	Rattus tunneyi
NC_014858.1	Rattus lutreolus
NC_005044.2	Capra hircus
NC_014770.1	Panthera tigris amoyensis
NC_014703.1	Cervus elaphus songaricus
NC_014701.1	Rucervus eldi
NC_014698.1	Pseudomys chapmani
NC_014696.1	Leggadina lakedownensis
NC_014692.1	Sus scrofa taiwanensis
NC_014456.1	Lynx rufus
NC_014453.1	Lepilemur hubbardorum
NC_014051.1	Nomascus siki
NC_014047.1	Symphalangus syndactylus
NC_014045.1	Hylobates pileatus
NC_014044.1	Bison bonasus
NC_014042.1	Hylobates agilis
NC_008415.3	Callorhinus ursinus
NC_007704.2	Cervus elaphus
NC_013993.1	Homo sp. Altai
NC_013996.1	Bos primigenius
NC_013840.1	Cervus elaphus yarkandensis
NC_013836.1	Cervus elaphus xanthopygus
NC_013834.1	Cervus nippon hortulorum
NC_013753.1	Moschus moschiferus
NC_013751.1	Naemorhedus caudatus
NC_013700.1	Nyctereutes procyonoides
NC_013571.1	Eothenomys chinensis
NC_013563.1	Proedromys liangshanensis
NC_013558.1	Vicugna vicugna
NC_013445.1	Cuon alpinus
NC_013276.1	Mesocricetus auratus
NC_013069.1	Budorcas taxicolor
NC_013068.1	Tscherskia triton
NC_012763.1	Loris tardigradus
NC_012775.1	Saimiri sciureus
NC_012774.1	Tarsius syrichta
NC_012773.1	Varecia variegata variegata
NC_012771.1	Eulemur macaco macaco
NC_012769.1	Eulemur fulvus mayottensis
NC_012766.1	Eulemur fulvus fulvus
NC_012764.1	Perodicticus potto
NC_012762.1	Otolemur crassicaudatus
NC_012761.1	Galago senegalensis
AC_000022.2	Rattus norvegicus strain Wistar
NC_007937.1	Balaenoptera omurai
NC_007936.1	Cricetulus griseus
NC_007938.1	Balaenoptera edeni
NC_007703.1	Rangifer tarandus
NC_007629.1	Lipotes vexillifer
NC_007441.1	Pantholops hodgsonii
NC_007393.1	Rousettus aegyptiacus
NC_007179.1	Cervus nippon yakushimae
NC_007009.1	Chlorocebus aethiops
NC_006993.1	Cervus nippon centralis
NC_006973.1	Cervus nippon yesoensis
NC_006931.1	Eubalaena japonica
NC_006928.1	Balaenoptera brydei
NC_006930.1	Eubalaena australis
NC_006925.1	Mystacina tuberculata
NC_006915.1	Mus musculus molossinus
NC_006926.1	Balaenoptera bonaerensis
NC_006924.1	Choloepus didactylus
NC_006927.1	Megaptera novaeangliae
NC_006929.1	Balaenoptera borealis
NC_006900.1	Trachypithecus obscurus
NC_006901.1	Colobus guereza
NC_006853.1	Bos taurus
NC_006835.1	Herpestes javanicus
NC_006295.1	Bubalus bubalis
NC_005971.1	Bos indicus
NC_005433.1	Rhinolophus monoceros
NC_005434.1	Rhinolophus pumilus
NC_005436.1	Pipistrellus abramus
NC_005435.1	Sorex unguiculatus
NC_005358.1	Ochotona princeps
NC_005314.1	Jaculus jaculus
NC_005315.1	Nannospalax ehrenbergi
NC_005275.1	Platanista minor
NC_005273.1	Hyperoodon ampullatus
NC_005274.1	Berardius bairdii
NC_005277.1	Pontoporia blainvillei
NC_005279.1	Monodon monoceros
NC_005272.1	Kogia breviceps
NC_005278.1	Lagenorhynchus albirostris
NC_005269.1	Caperea marginata
NC_005268.1	Balaena mysticetus
NC_005271.1	Balaenoptera acutorostrata
NC_005270.1	Eschrichtius robustus
NC_005280.1	Phocoena phocoena
NC_005276.1	Inia geoffrensis
NC_005212.1	Acinonyx jubatus

NC_002631.2	Echinops telfairi
NC_002503.2	Physeter catodon
NC_004029.2	Odobenus rosmarus rosmarus
NC_004030.2	Eumetopias jubatus
NC_005033.1	Hemiechinus auritus
NC_005034.1	Urotrichus talpoides
NC_005035.1	Mogera wogura
NC_004921.1	Elephantulus sp. VB001
NC_004920.1	Chrysochloris asiatica
NC_004919.1	Procavia capensis
NC_004577.1	Muntiacus crinifrons
NC_004563.1	Muntiacus muntjak
NC_002521.1	Tupaia belangeri
NC_002504.1	Lama pacos
NC_002369.1	Sciurus vulgaris
NC_002391.1	Talpa europaea
NC_000934.1	Loxodonta africana
NC_000889.1	Hippopotamus amphibius
NC_000884.1	Cavia porcellus
NC_002083.1	Pongo abelii
NC_002082.1	Hylobates lar
NC_002078.1	Orycteropus afer
NC_002009.1	Artibeus jamaicensis
NC_001992.1	Papio hamadryas
NC_001941.1	Ovis aries
NC_001913.1	Oryctolagus cuniculus
NC_001892.1	Glis glis
NC_001821.1	Dasypus novemcinctus
NC_001808.1	Ceratotherium simum
NC_001788.1	Equus asinus
NC_001779.1	Rhinoceros unicornis
NC_001700.1	Felis catus
NC_001646.1	Pongo pygmaeus
NC_001645.1	Gorilla gorilla
NC_001644.1	Pan paniscus
NC_001643.1	Pan troglodytes
NC_001640.1	Equus caballus
NC_001602.1	Halichoerus grypus
NC_001601.1	Balaenoptera musculus
NC_001325.1	Phoca vitulina
NC_001321.1	Balaenoptera physalus
NC_012706.1	Bos javanicus
NC_012694.1	Moschus berezovskii
NC_012684.1	Dicerorhinus sumatrensis
NC_012683.1	Rhinoceros sondaicus
NC_012682.1	Diceros bicornis
NC_012681.1	Coelodonta antiquitatis
NC_012670.1	Macaca fascicularis
NC_012461.1	Rattus praetor
NC_012389.1	Rattus exulans
NC_012387.1	Mus musculus castaneus
NC_012374.1	Rattus rattus
NC_012346.1	Bison bison
NC_012141.1	Martes flavigula
NC_011218.2	Canis lupus laniger
NC_012103.1	Pecari tajacu
NC_012102.1	Lama glama
NC_012100.1	Giraffa camelopardalis angolensis
NC_012098.1	Antilope cervicapra
NC_012096.1	Capricornis crispus
NC_012095.1	Sus scrofa domesticus
NC_012062.1	Grampus griseus
NC_012059.1	Tursiops truncatus
NC_012051.1	Stenella attenuata
NC_012061.1	Delphinus capensis
NC_012058.1	Tursiops aduncus
NC_012057.1	Sousa chinensis
NC_012053.1	Stenella coeruleoalba
NC_011822.1	Lama guanicoe
NC_011821.1	Hydropotes inermis
NC_011638.1	Rattus tanezumi
NC_011579.1	Martes zibellina
NC_008414.3	Rusa unicolor swinhoei
NC_011519.1	Macaca thibetana
NC_011358.1	Lutra lutra
NC_011304.1	Rhinolophus formosae
NC_011120.1	Gorilla gorilla gorilla
NC_011125.1	Meles meles
NC_011124.1	Ailurus fulgens
NC_011118.1	Ursus thibetanus thibetanus
NC_011117.1	Ursus thibetanus ussuricus
NC_011116.1	Arctodus simus
NC_011112.1	Ursus spelaeus
NC_011053.1	Propithecus coquereli
NC_011029.1	Ochotona curzoniae
NC_010650.1	Mus terricolor
NC_010640.1	Capricornis swinhoei
NC_010638.1	Uncia uncia
NC_010642.1	Panthera tigris
NC_010641.1	Panthera pardus
NC_010340.2	Canis lupus chanco
NC_010497.1	Spilogale putorius
NC_010339.1	Mus musculus musculus
NC_010304.1	Eremitalpa granti
NC_010298.1	Hylomys suillus
NC_010301.1	Dendrohyrax dorsalis
NC_010300.1	Eulemur mongoz
NC_010299.1	Daubentonia madagascariensis
NC_010302.1	Trichechus manatus
NC_009971.1	Ursus thibetanus
NC_009969.1	Tremarctos ornatus
NC_009970.1	Melursus ursinus
NC_009968.1	Helarctos malayanus
NC_009849.1	Camelus dromedarius
NC_009629.2	Camelus ferus
NC_009628.2	Camelus bactrianus
NC_009748.1	Chlorocebus tantalus
NC_009747.1	Chlorocebus pygerythrus
NC_009686.1	Canis lupus lupus
NC_009691.1	Ailurus fulgens styani
NC_009677.1	Meles anakuma
NC_009685.1	Gulo gulo
NC_009678.1	Martes melampus
NC_009692.1	Enhydra lutris
NC_009510.1	Ammotragus lervia
NC_009492.1	Ailuropoda melanoleuca
NC_006380.3	Bos grunniens
NC_009331.1	Ursus thibetanus formosanus
NC_002080.2	Erinaceus europaeus
NC_009126.1	Procyon lotor
NC_009056.1	Anomalurus sp. GP-2005
NC_008830.1	Phacochoerus africanus
NC_007596.2	Mammuthus primigenius
NC_008753.1	Ursus thibetanus mupinensis
NC_008749.1	Elaphodus cephalophus
NC_008491.1	Muntiacus reevesi micrurus
NC_008462.1	Cervus nippon taiouanus
NC_008450.1	Neofelis nebulosa
NC_008431.1	Pusa caspica
NC_008419.1	Neophoca cinerea
NC_008420.1	Arctocephalus townsendi
NC_008423.1	Lobodon carcinophaga
NC_008426.1	Erignathus barbatus
NC_008427.1	Cystophora cristata
NC_008430.1	Phoca largha
NC_008429.1	Phoca groenlandica
NC_008428.1	Phoca fasciata
NC_008417.1	Arctocephalus pusillus
NC_008418.1	Phocarctos hookeri
NC_008421.1	Monachus schauinslandi
NC_008425.1	Hydrurga leptonyx
NC_008432.1	Pusa sibirica
NC_008433.1	Pusa hispida
NC_008422.1	Mirounga leonina
NC_008424.1	Leptonychotes weddellii
NC_008416.1	Zalophus californianus
NC_008434.1	Vulpes vulpes
NC_001665.2	Rattus norvegicus strain BN/SsNHsdMCW
NC_008219.1	Piliocolobus badius
NC_008215.1	Semnopithecus entellus
NC_008217.1	Presbytis melalophos
NC_008218.1	Rhinopithecus roxellana
NC_008220.1	Pygathrix nemaeus
NC_008216.1	Nasalis larvatus
NC_008156.1	Galemys pyrenaicus
NC_008093.1	Canis latrans
NC_008092.1	Canis lupus
NC_008066.1	Chlorocebus sabaeus
NC_008064.1	Microtus levis
NC_005129.2	Elephas maximus
NC_004069.1	Muntiacus reevesi
NC_004028.1	Lepus europaeus
NC_004027.1	Manis tetradactyla
NC_004032.1	Tamandua tetradactyla
NC_004023.1	Arctocephalus forsteri
NC_004026.1	Macroscelides proboscideus
NC_004031.1	Galeopterus variegatus
NC_004025.1	Lemur catta
NC_003428.1	Ursus maritimus
NC_003427.1	Ursus arctos
NC_003426.1	Ursus americanus
NC_003314.1	Dugong dugon
NC_002008.4	Canis lupus familiaris
NC_003041.1	Microtus kikuchii
NC_003040.1	Episoriculus fumidus
NC_003033.1	Ochotona collaris
NC_002811.1	Tarsius bancanus
NC_002808.1	Echinosorex gymnura
NC_002764.1	Macaca sylvanus
NC_002763.1	Cebus albifrons
NC_002765.1	Nycticebus coucang
NC_002658.1	Thryonomys swinderianus
NC_002626.1	Chalinolobus tuberculatus
NC_002619.1	Pteropus scapulatus
NC_002612.1	Pteropus dasymallus
NC_018358.1	Elaphurus davidianus
NC_018116.1	Aotus nancymaae
NC_018115.1	Aotus azarai azarai

Appendix 4: Library quantification qPCR results and the first Herculase reamplification cycles.

SAMPLE	COPIES/µL AFTER LIBPREP	TOTAL COPIES AFTER LIBPREP	COPIES/µL AFTER INDEXING	TOTAL COPIES AFTER INDEXING	COPIES PER REAMP REACTION	AMP FOLD NEEDED TO E13	# CYCLES NEEDED	# REAMP CYCLES DONE
ZH0724	1.41E+06	5.22E+07	5.44E+08	2.67E+10	2.72E+09	3676	11	13
ZH0725	5.30E+05	1.96E+07	1.09E+08	5.35E+09	5.46E+08	18302	14	18
ZH0726	4.93E+06	1.82E+08	1.57E+09	7.68E+10	7.84E+09	1276	10	13
ZH0727	1.53E+06	5.64E+07	1.04E+08	5.09E+09	5.19E+08	19259	14	18
ZH0728	4.61E+06	1.70E+08	4.93E+08	2.42E+10	2.47E+09	4054	11	13
ZH0729	8.90E+05	3.29E+07	1.41E+08	6.91E+09	7.05E+08	14180	13	13
ZH0730	1.60E+07	5.90E+08	2.28E+09	1.12E+11	1.14E+10	879	9	9
ZH0731	2.17E+06	8.03E+07	3.30E+07	1.62E+09	1.65E+08	60624	15	18
ZH0732	1.65E+09	6.09E+10	6.63E+11	3.25E+13	3.31E+12	3	1	9
ZH0733	1.75E+08	6.48E+09	5.51E+10	2.70E+12	2.76E+11	36	5	9
ZH0734	1.10E+09	4.08E+10	6.44E+11	3.16E+13	3.22E+12	3	1	9
ZH0735	2.40E+08	8.87E+09	4.91E+10	2.40E+12	2.45E+11	41	5	9
ZH0736	7.81E+06	2.89E+08	2.35E+09	1.15E+11	1.17E+10	851	9	9
ZH0737	4.23E+05	1.57E+07	1.71E+08	8.39E+09	8.56E+08	11686	13	13
ZH0738	1.06E+07	3.91E+08	1.40E+09	6.85E+10	6.99E+09	1432	10	13
ZH0739	5.80E+07	2.15E+09	2.73E+10	1.34E+12	1.37E+11	73	6	9
ZH079EB1	1.15E+04	4.24E+05	4.62E+04	2.26E+06	2.31E+05	43299415	25	18
ZH079EB2	7.75E+03	2.87E+05	4.70E+06	2.30E+08	2.35E+07	425894	18	18
ZH079LB	5.97E+02	2.21E+04	1.15E+06	5.64E+07	5.76E+06	1736865	20	18

Appendix 5: Data quality control results

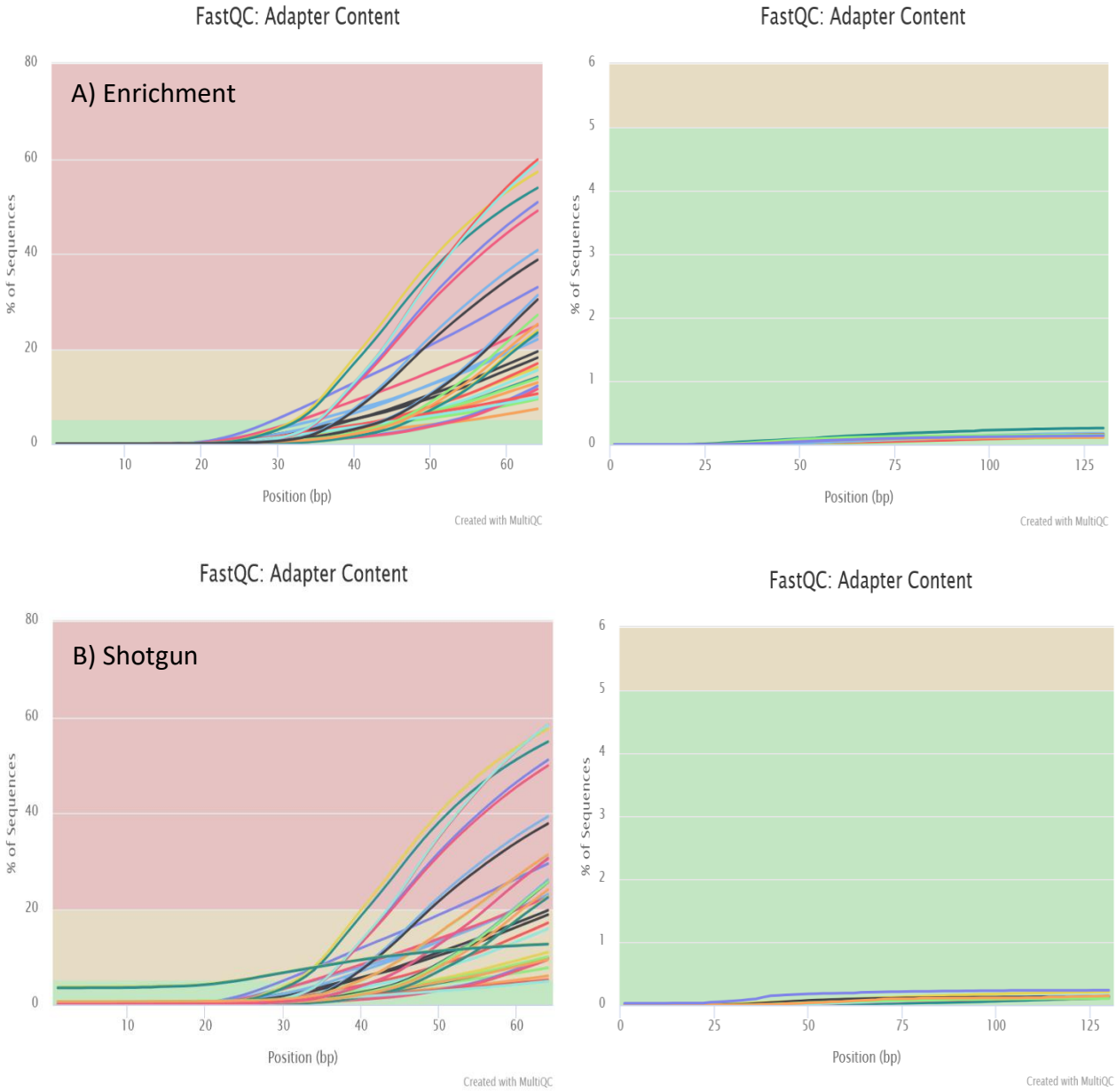

Figure A5.1: A) Adapter content in A) enriched and B) shotgun sequenced libraries before and after adapter removal and merging. Figures generated with MultiQC (Ewels et al., 2016).

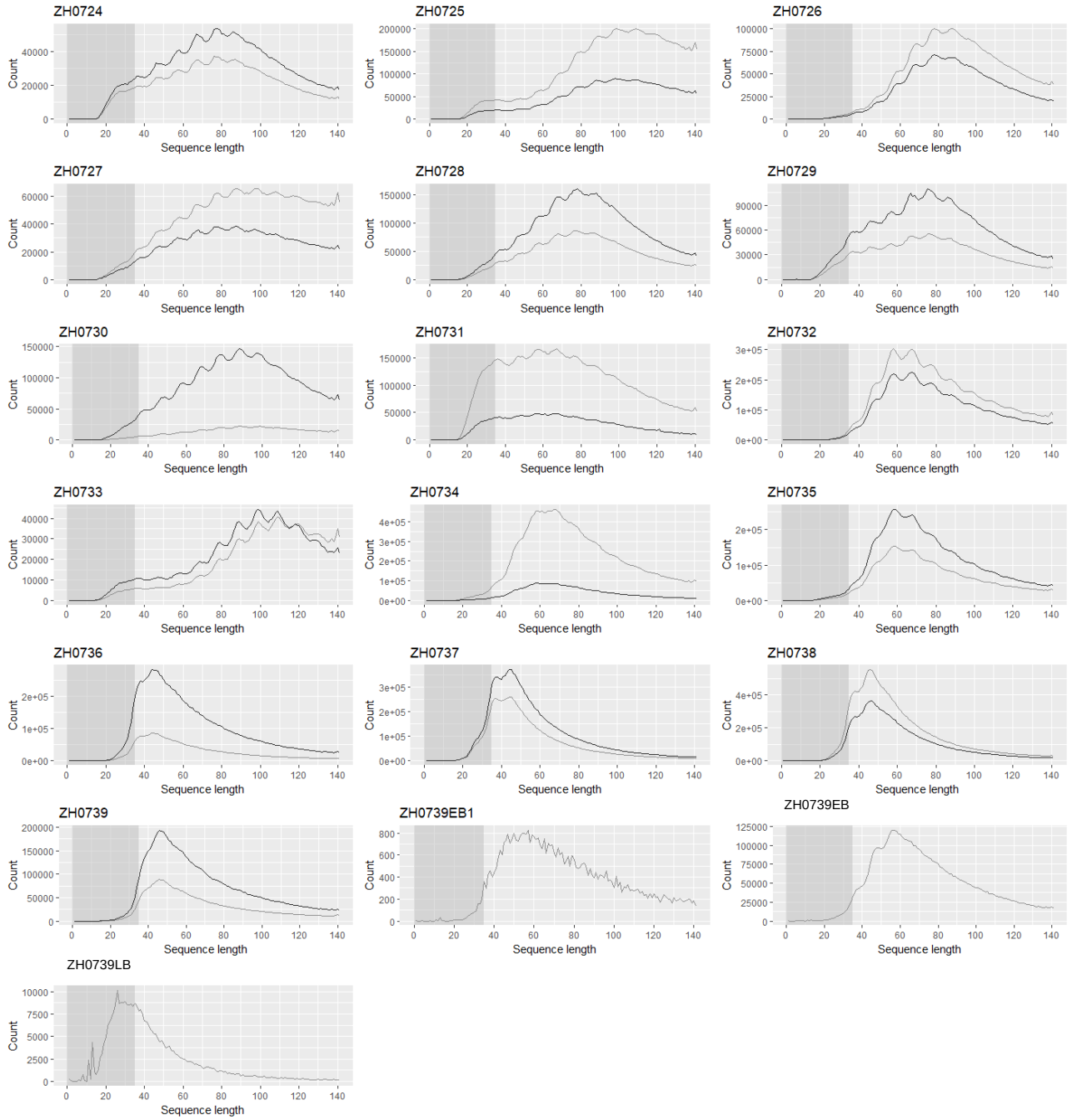

Figure A5.2: Fragment length distributions, reconstructed from merged sequences. Black lines indicate fragment length distributions in enriched libraries and grey lines in the corresponding shotgun sequenced libraries. The dark grey area shows the sequences below 35 bp, which were removed from the analysis.